# Analysis of Data Acquired Using ROC Paradigm and Its Extensions

**Dev P. Chakraborty**
University of Pittsburgh

**Xuetong Zhai**
University of Pittsburgh

## Abstract

A common task in medical imaging is assessing whether a new imaging system or device is an improvement over an existing one. Observer performance methodology, such as receiver operating characteristic (ROC) analysis, is widely used for this purpose and ROC studies are often required for regulatory approval of new devices. The purpose of this work is to describe **RJafroc**, which implements analysis of data acquired using the ROC paradigm and its location specific extensions. It is an enhanced implementation of existing Windows JAFROC (jackknife alternative free-response ROC, V4.2.1, http://www.devchakraborty.com) software. In the ROC paradigm the radiologist rates each image for confidence in presence of disease. In a common study design a number of radiologists (readers) rate images in two or more treatments (imaging systems), and the objective is to determine the significances of the inter-treatment differences between reader-averaged FOMs. In the free-response ROC paradigm the reader marks the locations of suspicious regions and rates each region for confidence in presence of disease, and credit for detection is only given if a true lesion is correctly localized. In the region of interest (ROI) paradigm each image is divided into a number of ROIs and the reader rates each ROI. Each paradigm requires definition of a valid FOM that rewards correct decisions and penalizes incorrect ones and specialized significance testing procedure are applied. The package reads data in all currently used data formats including Excel. Significance testing uses two models in widespread use. Included are tools for (1) calculating a variety of free-response FOMs; (2) ROC sample size estimation for planning a future study based on pilot data; (3) viewing empirical operating characteristics in ROC and free-response paradigms; (4) producing formatted report files; and (5) saving data files in appropriate formats for analysis with alternate software. In addition to open-source access to the functions, the package includes a graphical interface for users already familiar with the Windows software, who simply wish to run the program.

*Keywords*: medical imaging, observer performance, assessment methodology, ROC, FROC, JAFROC software, R.

# 1. Introduction

A common task in medical imaging is assessing whether a new imaging system is an improvement over an existing one. Observer performance measurements, widely used for this purpose, require data collection and analyses methods that fall under the rubric of what is loosely termed "ROC analysis", where ROC is an abbreviation for Receiver Operating Characteristic (Metz 1986). ROC analysis is a specialized branch of statistics that is important in diagnostic imaging, where new imaging technologies and the accuracy of radiologist interpretations need to be assessed. The Food and Drug Administration (FDA), which regulates medical imaging devices, requires ROC studies as part of the device approval process (see document "Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests" available at http://www.fda.gov/RegulatoryInformation/Guidances). There are, conservatively, at least 1000 publications describing medical imaging ROC studies and a seminal paper (Metz 1986) by the late Prof. C.E. Metz has been cited over 1800 times. Since they involve numbers of radiologists interpreting large number of images in different modalities, ROC studies can be expensive to conduct. For example (Pisano, Gatsonis, Hendrick, Yaffe, Baum, Acharyya, Conant, Fajardo, Bassett, D'Orsi, Jong, and Rebner 2005), the Digital Mammography for Imaging Screening Trial (DMIST) cost about $30 million (this study involved about 50,000 asymptomatic women at 33 mammography centers, and each mammogram was interpreted by two radiologists per mammography center). More typical ROC studies proposed in National Institutes of Health (NIH) grant applications are budgeted in the hundreds of thousands of dollars and often take years to complete. Consequently, there is much interest in optimizing methodology for analyzing ROC studies and its extensions, and four websites currently disseminate software for analyzing such studies: the University of Chicago has a site for ROC analysis software (http://metz-roc.uchicago.edu/) as does the University of Iowa (http://perception.radiology.uiowa.edu/) and the FDA (https://code.google.com/p/imrmc/); Windows software called JAFROC (jackknife alternative free-response ROC) (Chakraborty and Zhai 2014), which can analyze ROC studies and its extensions (Chakraborty and Berbaum 2004; Chakraborty 2013), is available at http://www.devchakraborty.com. Software from the University of Iowa and University of Chicago websites have been used in several hundred publications (Professor Kevin Berbaum, University of Iowa, personal communication, ca 2014). JAFROC has been used in 78 publications: the list is viewable at http://www.devchakraborty.com/JafrocApplications.pdf. The purpose of this work is to describe **RJafroc** Version 0.1.1, which is an enhanced implementation of JAFROC.[1] In this section we introduce terminology used in the `RJafroc-package` page of the documentation that accompanies this paper. For details regarding basics of ROC methodology, reviews of this field may be consulted (Metz 1978, 1986, 1989; Wagner, Beiden, Campbell, Metz, and Sacks 2002; Wagner, Metz, and Campbell 2007; Kundel, Berbaum, Dorfman, Gur, Metz, and Swensson 2008; Metz 2008).

In an ROC study the radiologist is shown images of *patients* (*images* and *cases* are used interchangeably as synonyms for patients), the radiologist is "blinded", of course, to the true disease states, and the radiologist's task is to *rate* each patient for confidence in presence or absence of disease. The rating $r$ is typically on an ordinal scale, with higher values representing increasing confidence in presence of disease. Typically 5 or 6 integer ratings are used but the ratings could have higher precision (Metz, Herman, and Shen 1998b). With a 6 rating scale a 1-rating would correspond to high confidence that patient is non-diseased and a 6-rating would correspond to high confidence that patient is diseased. The normalized counts in the different ratings bins, cumulated separately for actually non-diseased and actually diseased patients, can be used to construct an *operating point*. For example, the cumulated

[1] A number of ROC packages are currently available on CRAN and Bioconductor: **ROCR** (Sing, Sander, Beerenwinkel, and Lengauer 2005), **pROC** (Robin, Turck, Hainard, Tiberti, Lisacek, Sanchez, and Müller 2011), **ROCwoGS** (Wang 2012), **ROCt** (Foucher, Trebern-Launay, and Lorent 2013) and **ROC632** (Foucher 2013). These are intended for classifier performance evaluation and visualization, in the machine learning, pattern recognition, artificial intelligence and genetics areas. None of them relate to medical imaging applications addressed in the software available on the 4 previously mentioned websites.

counts in diseased ratings bins 3, 4 and 5, divided by the number of actually diseased images, yields *true positive fraction* $TPF_{3+}$, where $TPF$ is the ordinate of the ROC plot, and the corresponding cumulated counts for non-diseased images, divided by the number of non-diseased images, yields *false positive fraction* $FPF_{3+}$, where $FPF$ is the abscissa of the ROC plot. True positive fraction is synonymous with *sensitivity* and the complement of false positive fraction is synonymous with *specificity*, so the ROC curve is a plot of *sensitivity* vs. $1 - specificity$. It can be seen that as long as no bin has zero counts for both non-diseased and diseased images, an $R$ rating ROC study will yield $R - 1$ non-trivial operating points $\{FPF_{r+}, TPF_{r+}; r = 2, 3, ..., R\}$. The origin $(0, 0)$ and the upper right corner $(1, 1)$ are trivial operating points, belonging to any ROC dataset, obtained by counting none and all of the binned ratings, respectively. The empirical ROC curve is defined by connecting neighboring operating points (including the trivial ones) with straight lines. While several curve-fitting methods are available (Dorfman and Alf 1968, 1969; Dorfman and Berbaum 1986; Dorfman, Berbaum, Metz, Lenth, Hanley, and Abu Dagga 1997; Pan and Metz 1997; Metz and Pan 1999; Dorfman and Berbaum 2000; Pesce and Metz 2007) and have their merits, the trapezoidal area under the empirical ROC is frequently used as a non-parametric FOM for quantifying observer performance (Hanley and McNeil 1982). It can be shown to be equivalent to the Mann-Whitney-Wilcoxon 2-sample U-statistic (Wilcoxon 1945; Mann and Whitney 1947). ROC studies are typically conducted with about 50/50 or more non-diseased/diseased patients. The patients are imaged in two or more imaging systems (termed *modalities* or *treatments*) and the images are rated by a number of radiologists (typically about 5 to 10). This type of fully crossed study design is termed multiple reader multiple case (MRMC) and, although methods are available for partially paired interpretations (Metz, Herman, and Roe 1998a; Obuchowski 2009), MRMC studies are the focus of this work.

A limitation of the ROC paradigm is that it acquires a single rating per image, where the rating applies to the image as a whole, not to any specific region(s) in the image. Typically, disease is manifested by the presence of localized diseased regions or *lesions*. For example, lung cancer often presents as localized malignant nodules found on chest x-rays or computed tomography (CT) scans of the chest. Ignoring localization can result in an overestimate of true performance (Obuchowski, Mazzone, and Dachman 2010); for example, suppose a true lesion on a diseased case is missed and a disease-free region is perceived as abnormal by the radiologist - the two mistakes would effectively cancel each other and the event would be credited as a true positive at the level of confidence associated with the disease-free region. This and othe ambiguities (Bunch, Hamilton, Sanderson, and Simmons 1978) associated with neglect of location information has been the reason for research on extending ROC analysis to location-specific interpretations.

There are two data collection paradigms that allow for localization information to be collected to different extents (a third important paradigm (Starr, Metz, Lusted, and Goodenough 1975; Starr, Metz, and Lusted 1977; Swensson and Judy 1981; Swensson 1996), termed *location ROC* (LROC) is not included in this description, as it is not currently implemented in any of the websites mentioned previously). In the *free-response* paradigm (Egan, Greenburg, and Schulman 1961; Miller 1969; Bunch *et al.* 1978) the radiologist *marks* and *rates* regions that are suspicious for disease. A mark is classified as *lesion localization* (LL) if it successfully locates an actual lesion to within clinically acceptable spatial accuracy, or *non-lesion localization* (NL) otherwise (usage of ROC-specific terms like true positive and false positive in the FROC (free-response ROC), LROC or ROI contexts can lead to confusion). Unmarked lesions are assigned the –infinity rating. By treating the rating of the highest rated mark on a *non-diseased* image (or –infinity if the image has no marks) as its *inferred* FP rating, it is possible to define an inferred FPF quantity that is analogous to true FPF obtained in an actual ROC study. By cumulating LL events and dividing by the total number of lesions it is possible to define a *lesion localization fraction* (LLF) quantity that is analogous to TPF, but because it requires correct localization, may not reach unity, even when all ratings are cumulated. A plot of LLF along the ordinate vs. FPF is defined as the *alternative* FROC, or AFROC (Chakraborty 1989; Chakraborty and Winter 1990), where it is understood that the uppermost operating point, obtained by cumulating all the marks, is to be connected to (1,1) by a dotted line. While inaccessible to the observer, the area under the dotted line needs to be taken into account in the definition of the area under the AFROC as a valid FOM (Chakraborty 2006b,a); essentially it gives credit for unmarked non-diseased cases and penalizes for unmarked lesions). Non-lesion localization fraction (NLF) is defined as the cumulated number of NLs divided by the total number of cases. The FROC plot is defined as that of LLF along the ordinate vs. NLF (Bunch, Hamilton, Sanderson, and Simmons 1978; Chakraborty, Breatnach, Yester, Soto, Barnes, and Fraser 1986; Niklason, Hickey, Chakraborty, Sabbagh, Yester, Fraser, and Barnes 1986; Barnes, Sabbagth, Chakraborty, Nath, Luna, Sanders, and Fraser 1989). By treating the rating of the highest rated mark on a diseased image (or –infinity if the image has no marks) as its inferred TP rating, it is possible to define an inferred TPF. The plot of inferred TPF vs. inferred FPF is the inferred ROC curve. Regarding the highest rated NL mark on *any* image as an inferred FP1 rating (the 1 denotes that NL marks on diseased cases could be contributing to this FP-like rating) and the corresponding AFROC1 plot is that of LLF vs. FPF1. By assigning clinically relevant weights to different lesions on the same diseased image, it is possible to define weighted LLF, weighted AFROC and weighted AFROC1 plots (the weights, which add up to unity on any diseased image, are the relative importances of finding the lesions: from the clinical perspective all lesions are not alike; some are more aggressive than others and therefore more important to find at an early stage). With the exception of the FROC, the trapezoidal areas under all of these curves qualify as valid FOMs (a valid FOM is one that rewards good decisions and penalizes bad decisions, where good and bad are defined with respect to patient outcome). That the area under the FROC is a particularly poor FOM can be appreciated from the fact that a perfect observer's FROC curve would be a vertical line extending from (0,0) to (0,1), for which the area measure would be zero.

In the *region of interest* (ROI) paradigm (Obuchowski, Lieber, and Powell 2000) each image is divided into Q regions of interest (typically Q is about 4) where each region is either non-diseased or diseased, and the reader gives an ROC-like rating to each region. Regarding each of the regions as a mini-image, it is possible to define ROC-like quantities TPF' and FPF', where the primes distinguish them from true FPF and TPF. For example, FPF' and TPF' can be defined for a dataset containing only diseased images, for which it would be impossible to define FPF. The data collection paradigms are summarized in Table 1.

Analysis of the data starts with estimation, for each treatment - reader combination, of the FOM. One object of the analysis is to determine the significance of the reader-averaged differences in FOMs between pairs of modalities. While several significance-testing methods have been proposed, see Table 2, we focus on two that are easily accessible and consequently in widespread use: the *Dorfman-Berbaum-Metz* (DBM) method (Dorfman, Berbaum, and Metz 1992) and the *Obuchowski-Rockette* (OR) method (Obuchowski and Rockette 1995), both of

| Data collection paradigm | Operating characteristic(s) | FOM | Terminology |
|---|---|---|---|
| Receiver operating characteristic | ROC = TPF vs. FPF | Wilcoxon | AUC |
| Free-response | AFROC = LLF vs. FPF | Trapezoidal area under AFROC | JAFROC, weighted JAFROC |
| | AFROC1 = LLF vs. FPF1 | Trapezoidal area under AFROC1 | JAFROC1, weighted JAFROC1 |
| | FROC = LLF vs. NLF | Not recommended | |
| | Inferred ROC | Trapezoidal area under inferred ROC | AUC |
| Region of interest | ROC'=TPF' vs. FPF' | Trapezoidal area under ROC' | AUC' |

Table 1: Data collection, paradigms and associated operating characteristics, FOMs and terminology. [AUC = area under the ROC curve, plot of TPF vs. FPF; AUC' = area under the ROC' curve, plot of TPF' vs. FPF']

which have been significantly improved by contributions by Hillis, and are henceforth referred to as DBMH and ORH, respectively. A third method (Gallas 2006; Gallas, Bandos, Samuelson, and Wagner 2009) often termed a mechanistic or first-principles approach to MRMC analysis, is also available online, that yields independent estimates of variability parameters used in DBMH and ORH analyses, in addition to its own estimates. *All significance-testing methods are applicable to any scalar FOM.*

If a non-significant result is obtained (i.e., $p > \alpha$ ) in a *pilot* study then the investigator may wish to plan a *pivotal* study that is sufficiently powered to detect a clinically relevant difference in FOMs between the two modalities of interest. The pilot study is used to get estimates of variability components in the analysis model, and the approximate expected effect size. Sample-size estimation methods for ROC studies are available on all referenced websites. A preliminary sample-size method for free-response studies is available on the JAFROC website. We are unaware of any sample size estimation method for ROI studies.

## 2. Statistical models and methods

The FOM is a critical determinant of statistical power (Chakraborty 2008) and clinical relevance (Chakraborty 2012) of the measurement. Even for the relatively simple ROC paradigm, several FOMs have been proposed, e.g., partial area measures (Jiang, Metz, and Nishikawa 1996; Yousef, Wagner, and Loew 2005), the Youden index (Youden 1950) and others (Pepe 2003). In the following sections we define the implemented ROC data FOM, two FOMs commonly used in analyzing free-response data (several other implemented free-response FOMs

| Significance testing methods | Online software name and website | Supported data collection paradigms | Supported FOMs |
|---|---|---|---|
| DBMH, ORH | OR-DBM MRMC (Hillis, Schartz, and Berbaum 2015) | ROC | Wilcoxon and parametric fits derived AUCs, and others |
| DBMH | Metz ROC Software (Pesce, Papaioannou, and Metz 2011) | | Wilcoxon and parametric fits derived AUCs, and others |
| Mechanistic MRMC | iMRMC (Gallas, He, and Pathare 2015) | | Wilcoxon |
| DBMH | JAFROC (Chakraborty and Zhai 2014) | ROC, FROC, ROI | Trapezoidal areas under ROC, AFROC, AFROC1, weighted versions and trapezoidal area under ROC', and other FOMs |
| Ordinal regression(Toledano and Gatsonis 1996; Toledano 2003) | NA | ROC | Wilcoxon and others |
| Wald test on U-statistics(Song 1997) | | | |
| Hierarchical ordinal regression(Ishwaran and Gatsonis 2000; Obuchowski, Beiden, Berbaum, Hillis, Ishwaran, Song, and Wagner 2004) | | | |
| Multiple bootstraps(Beiden, Wagner, and Campbell 2000) | | | |

Table 2: Software availability of MRMC observer performance methods.

are defined in Appendix A.1), followed by the ROI FOM. Two implemented significance-testing methods are described followed by sample-size estimation for ROC studies. No derivations are given: we simply refer the interested reader to the literature.

## 2.1. Notation and FOM for ROC data

Images are indexed by $k_t t$, where $t$ is the truth state (1 for disease-free cases and 2 for diseased cases) and $k_t$ indexes the cases for truth state $t$, specifically, $k_1 = 1, 2 \ldots, K_1$ and $k_2 = 1, 2 \ldots, K_2$ where $K_1$ is the number of disease-free cases and $K_2$ is the number of diseased cases. Let $r_{ijk_t t}$ denote the rating given to case $k_t t$ by reader $j$ using modality $i$, where $i = 1, 2 \ldots, I$ and $j = 1, 2 \ldots, J$; $I$ is the number of modalities and $J$ is the number of readers. The trapezoidal area under the ROC curve, $\theta$, estimated for reader $j$ in modality $i$ by the Wilcoxon statistic (Wilcoxon 1945; Mann and Whitney 1947):

$$\widehat{\theta}_{ij} = \frac{1}{K_1 K_2} \sum_{k_1}^{K_1} \sum_{k_2}^{K_2} \psi\left(r_{ijk_1 1}, r_{ijk_2 2}\right) \tag{1}$$

The kernel function $\psi$ is defined by:

$$\left.\begin{array}{ll} \psi\left(r_{ijk_1 1}, r_{ijk_2 2}\right) = 1 & r_{ijk_1 1} < r_{ijk_2 2} \\ \psi\left(r_{ijk_1 1}, r_{ijk_2 2}\right) = 0.5 & r_{ijk_1 1} = r_{ijk_2 2} \\ \psi\left(r_{ijk_1 1}, r_{ijk_2 2}\right) = 0 & r_{ijk_1 1} > r_{ijk_2 2} \end{array}\right\} \tag{2}$$

This FOM is identical to the area under the empirical (trapezoidal) ROC curve (Bamber 1975). It has the physical interpretation as the probability that a randomly picked diseased image will be rated higher than a randomly picked non-diseased image (Hanley and McNeil 1982).

## 2.2. Notation and FOMs for free-response data

Since free-response data allows for varying number of lesions and mark/rating pairs per case, the notation is necessarily more complex. The *case-truth* index $t$ refers to the case (or patient) as a whole (non-diseased, $t = 1$, or diseased, $t = 2$), not to specific locations in the case. Let $N_{k_2 2}$ denote the number of lesions in diseased case $k_2 2$, where $N_{k_2 2} \geq 1$. The total number of lesions in the data set is $N_2$:

$$N_2 = \sum_{k_2=1}^{K_2} N_{k_2 2} \tag{3}$$

The notation is drived from the Chakraborty *search-model* for the free-response paradigm (Chakraborty 2006a,b) that involves two phases, a *search phase* during which suspicious regions (*decision sites*) are identified (based on eye-tracking measurements on radiologists this phase is quite rapid (Kundel, Nodine, Conant, and Weinstein 2007), typically 100 ms for experts) and a *decision phase* during which each decision-site is examined (typically 1 sec per site) and a decision is made on whether to mark it. Decision sites can be either *noise sites* (not corresponding to real lesions) or *signal sites* (corresponding to real lesions). Marked noise sites are non-lesion localizations while marked signal sites are lesion localizations. Marks are labeled by a *location index* $l_s$ $(l_s = 1, 2, \ldots)$ *and* a *site-truth* index $s$ which determines the *type* of the site, i.e., $s = 1$ for a non-lesion localization and $s = 2$ for a lesion localization. The rating for modality $i$, reader $j$, case $k_t t$ and site $l_s s$ is denoted $r_{ijk_t t l_s s}$.

Several methods have been proposed to infer ROC-like data (i.e., single rating per image) from free-response data. The highest rating inferred ROC (IR) FOM $\theta_{ij}^{IR}$ is estimated by (this is identical to the A0 FOM defined by Song, Bandos, Rockette, and Gur (2008)):

$$\widehat{\theta}_{ij}^{IR} = \frac{1}{K_2 K_1} \sum_{k_2=1}^{K_2} \sum_{k_1=1}^{K_1} \psi \left( max(r_{ijk_1 1 * 1}), max(r_{ijk_2 2 * *}) \right) \tag{4}$$

The *max* function is the maximum over the indices indicated by the asterisks. For the second max function the maximum is over all marks (NLs and LLs) on the diseased case. If all lesions are marked and no noise sites are marked, signifying perfect performance, the $\psi$ function is unity, and $\widehat{\theta}_{ij}^{IR}$ is unity. If no lesions are marked and the distribution of the numbers and ratings of NL marks is the same for non-diseased and diseased images, signifying the observer is unable to discriminate between them, the $\psi$ function comparisons yield 0.5, on the average, implying $\widehat{\theta}_{ij}^{IR} = 0.5$, which is the worst possible ROC performance. The Song et al (Song *et al.* 2008) A1 figure of merit takes the average rating of all marked regions on an image to infer an ROC-like rating for the image. The Song A2 FOM involves a stochastic dominance idea described in (Song *et al.* 2008).

Let $W_{k_2 l_2}$ denote the weight (clinical importance) of lesion $l_2 2$ in abnormal case $k_2 2$ such the weights on any given diseased case add up to unity:

$$\sum_{l_2=1}^{N_{k_2 2}} W_{k_2 l_2} = 1 \tag{5}$$

The weighted (according to clinical importance) JAFROC FOM $\theta_{ij}^{wJAFROC}$ is estimated by:

$$\widehat{\theta}_{ij}^{wJAFROC} = \frac{1}{K_2 K_1} \sum_{k_2=1}^{K_2} \sum_{k_1=1}^{K_1} \sum_{l_2=1}^{N_{k_2 2}} W_{k_2 l_2} \psi \left( max(r_{ijk_1 1 * 1}), r_{ijk_2 2 l_2 2} \right) \tag{6}$$

If all lesions are marked and no non-diseased image is marked the $\psi$ function is unity and $\widehat{\theta}_{ij}^{wJAFROC}$ is unity (the best possible performance). If no lesions are marked and every non-diseased image has at least one mark the $\psi$ function is zero and $\widehat{\theta}_{ij}^{wJAFROC}$ is zero (the worst possible performance). This figure or merit, like the one to be described next, ranges between 0 and unity, unlike the ROC area FOM that ranges between 0.5 and 1. The above FOM does not count NLs on diseased cases. The extension to include the highest rated NL on diseased cases, called the weighted JAFROC1 FOM, $\widehat{\theta}_{ij}^{wJAFROC1}$, is:

$$\begin{aligned} \widehat{\theta}_{ij}^{wJAFROC1} = \frac{1}{K_2 \left( K_1 + K_2 \right)} \sum_{k_2=1}^{K_2} & \left[ \sum_{k_1=1}^{K_1} \sum_{l_2=1}^{N_{k_2 2}} W_{k_2 l_2} \psi \left( max(r_{ijk_1 1 * 1}), r_{ijk_2 2 l_2 2} \right) \right. \\ & \left. + \sum_{k_2'=1}^{K_2} \sum_{l_2=1}^{N_{k_2 2}} W_{k_2 l_2} \psi \Big( max(r_{ijk_2' 2 * 1}), r_{ijk_2 2 l_2 2} \Big) \right] \end{aligned} \tag{7}$$

The first term in the numerator compares LL ratings to the maximum NL ratings on non-diseased images, similar to Eqn. 6. The second term compares LL ratings to the maximum NL ratings on diseased images. Since the maximum of NL ratings in $k_2'2$ is being compared with each LL rating in $k_22$, we should use the lesion weights corresponding to $k_22$ and the $l_2$ index ranges from 1 to $N_{k_22}$. The above two FOMs have covered the needs of most users of JAFROC. Other implemented free-response FOMs, sometimes needed for specific clinical reasons, are described in Appendix A.1.

## 2.3. Notation and FOM for ROI data

In this paradigm each image is divided into $Q_{k_tt}$ regions of interest (ROIs). Obuchowski's analytic significance testing procedure (Obuchowski 1997) can handle varying number of ROIs per image, but is currently unimplemented in **RJafroc**, which instead uses resampling methods for significance testing. Let $r_{ijk_22l_22}$ denote the rating in modality $i$, reader $j$, for the lesion-present ROI indexed by $l_22$ in diseased case $k_22$ and let $q_{k_222}$ denote the total number of lesion-containing ROIs in the case. Similarly, let $r_{ijk_ttl_11}$ denote the rating in modality $i$, reader $j$, for the lesion-absent ROI indexed by $l_11$ in case $k_tt$ (which could be non-diseased or diseased) and let $q_{k_tt1}$ be the total number of non-lesion containing ROIs in the case. The trapezoidal area under the ROI-level ROC curve is estimated by Obuchowski *et al.* (2000)

$$\widehat{\theta}_{ij}^{ROI} = \frac{\sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2}\sum_{t=1}^{2}\sum_{l_1=1}^{q_{k_tt1}}\sum_{l_2=1}^{q_{k_222}} \psi\left(r_{ijk_ttl_11}, r_{ijk_22l_22}\right)}{\sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2}\sum_{t=1}^{2}\sum_{l_1=1}^{q_{k_tt1}}\sum_{l_2=1}^{q_{k_222}} (1)} \tag{8}$$

For $t = 1$ the comparisons are between ratings of lesion-containing ROIs and ratings of ROIs on non-diseased cases and for t = 2 comparisons are between ratings of lesion-containing ROIs and ratings of lesion-absent ROIs on diseased cases. Unlike the ROC FOM and the weighted JAFROC FOM, the ROI FOM, like the weighted JAFROC1 FOM, can be defined over a dataset with no non-diseased cases. Table 3 summarizes the FOMs described so far.

## 2.4. DBMH significance testing method

The DBM method (Dorfman *et al.* 1992) models the jackknife derived pseudovalues (Efron and Tibshirani 1993) of $\widehat{\theta}_{ij}$, denoted $Y'_{ijk}$ for modality $i$, reader $j$ and case $k$ ($k = 1, 2, \ldots K$; where $K = K_1 + K_2$ is the total number of cases). The pseudovalues are defined by:

$$Y'_{ijk} = K\widehat{\theta}_{ij} - (K-1)\,\widehat{\theta}_{ij(k)} \tag{9}$$

Here $\widehat{\theta}_{ij(k)}$ is the estimate of $\theta_{ij}$ for modality $i$, reader $j$ and case $k$ removed (jackknifed) from the analysis. Hillis, Berbaum, and Metz (2008) have defined a centering transformation

$$Y_{ijk} = Y'_{ijk} + \left(\widehat{\theta}_{ij} - Y'_{ij\bullet}\right) \tag{10}$$

The effect of this transformation is that the average of the centered pseudovalues over the

| Paradigm | Description of FOM | Symbol | Comments |
|---|---|---|---|
| ROC | Trapezoidal area under ROC | $\widehat{\theta}_{ij}$ | Equivalent to Wilcoxon statistic |
| FROC | Highest rating inferred ROC | $\widehat{\theta}_{ij}^{IR}$ | |
| | Average rating inferred ROC | A1 | |
| | Stochastic dominance inferred ROC | A2 | |
| | Weighted JAFROC | $\widehat{\theta}_{ij}^{wJAFROC}$ | Recommended FOM for FROC data |
| | Weighted JAFROC1 | $\widehat{\theta}_{ij}^{wJAFROC1}$ | To be used only in absence of non-diseased cases |
| ROI | Trapezoidal area under ROI-level ROC' | $\widehat{\theta}_{ij}^{ROI}$ | |

Table 3: Summary of the FOMs for the different observer performance measurement data collection methods. [IR = inferred ROC using the highest rating; A1, A2 are inferred ROC FOMs]

case index is identical to the estimate of the FOM:

$$Y_{ij\bullet} = Y'_{ij\bullet} + \left(\widehat{\theta}_{ij} - Y'_{ij\bullet}\right) = \widehat{\theta}_{ij} \tag{11}$$

This has the practical advantage that all confidence intervals are correctly centered. While this transformation is unnecessary if one uses the Wilcoxon as the figure-of-merit, for generality with other possible FOMs, *it is understood that all calculations from now on will use centered pseudovalues*. The DBM pseudovalue model (Dorfman *et al.* 1992) is:

$$\begin{aligned} &Y_{ijk} = \mu + \tau_i + R_j + C_k + (\tau R)_{ij} + (\tau C)_{ik} + (RC)_{jk} + \varepsilon_{ijk} \\ &\sum_{i=1}^{I} \tau_i = 0 \end{aligned} \tag{12}$$

The right hand side consists of 2 fixed effects, $\mu, \tau_i$, and 6 random effects modeled as mutually independent samples from zero-mean normal distributions with variances (in the same order of appearance in the above equation) $\sigma_R^2$, $\sigma_C^2$, $\sigma_{\tau R}^2$, $\sigma_{\tau C}^2$, $\sigma_{RC}^2$ and $\sigma_{\varepsilon}^2$. Using the dot symbol to denote an average over the corresponding index, the first term can be $\mu$ estimated by averaging the observed left hand side over all three indices:

$$\mu = Y_{\bullet\bullet\bullet} \tag{13}$$

The modality effect can be estimated by:

$$\tau_i = Y_{i\bullet\bullet} - \mu \tag{14}$$

The reader and case averaged difference between two different modalities $i$ and $i'$ (often termed the *observed effect size*) is given by

$$\tau_i - \tau_{i'} = Y_{i\bullet\bullet} - Y_{i'\bullet\bullet} = \widehat{\theta}_{i\bullet} - \widehat{\theta}_{i'\bullet} \tag{15}$$

Estimating the strengths of the random terms involves analysis of variance (ANOVA) methods specially adapted to this problem by Dorfman, Berbaum, Metz, Hillis and others. The starting point is calculation of the mean squares. In the following definitions the Y subscript emphasizes that the relevant mean-square quantities are calculated using pseudovalues, not figure-of-merit values.

$$\begin{aligned}
MS_Y(T) &= \frac{JK\sum\limits_{i=1}^{I}(Y_{i\bullet\bullet}-Y_{\bullet\bullet\bullet})^2}{I-1}\\
MS_Y(R) &= \frac{IK\sum\limits_{j=1}^{J}(Y_{\bullet j\bullet}-Y_{\bullet\bullet\bullet})^2}{J-1}\\
MS_Y(TR) &= \frac{K\sum\limits_{i=1}^{I}\sum\limits_{j=1}^{J}(Y_{ij\bullet}-Y_{ij\bullet}-Y_{\bullet j\bullet}+Y_{\bullet\bullet\bullet})^2}{(I-1)(J-1)}\\
MS_Y(TC) &= \frac{J\sum\limits_{i=1}^{I}\sum\limits_{k=1}^{K}(Y_{i\bullet k}-Y_{i\bullet\bullet}-Y_{\bullet\bullet k}+Y_{\bullet\bullet\bullet})^2}{(I-1)(K-1)}\\
MS_Y(\varepsilon) &= \frac{\sum\limits_{i=1}^{I}\sum\limits_{j=1}^{J}\sum\limits_{k=1}^{K}\left(Y_{ijk}-Y_{ij\bullet}-Y_{i\bullet k}-Y_{\bullet jk}+Y_{i\bullet\bullet}+Y_{\bullet j\bullet}+Y_{\bullet\bullet k}-Y_{\bullet\bullet\bullet}\right)^2}{(I-1)(J-1)(K-1)}
\end{aligned} \tag{16}$$

Hillis proposes the following $f$ statistic for testing the null hypothesis of no modality effect (Hillis 2007):

$$F_{DBMH} = \frac{MS_Y(T)}{MS_Y(TR) + H\left(MS_Y(TC) - MS_Y(\varepsilon)\right)} \tag{17}$$

Here $H(x)$ is the unit step function, defined as unity for positive $x$ and zero otherwise. Hillis has shown that $F_{DBMH}$ is distributed as an $f$ statistic with numerator degrees of freedom $ndf = I - 1$ (i.e., one less than the number of treatments) and $ddf_H$ denominator degrees of freedom, i.e.,

$$F_{DBMH} \sim F_{ndf,ddf_H} \tag{18}$$

The denominator degrees of freedom $ddf_H$ is defined by (this is different from the original definitions by DBM):

$$ddf_H = \frac{\left[MS_Y(TR) + H\left[MS_Y(TC) - MS_Y(\varepsilon)\right]\right]^2}{\frac{MS_Y(TR)^2}{(I-1)(J-1)}} \tag{19}$$

The critical value of the $f$ statistic for rejection of the null hypothesis is given by $F_{1-\alpha,ndf,ddf_H}$. The $p$ value of the test is given by:

$$p = P\left(F > F_{DBMH} | F \sim F_{ndf,ddf_H}\right) \tag{20}$$

The $(1-\alpha)\,100$ percent confidence interval for $(\theta_i - \theta_{i'})$is given by

$$CI_{1-\alpha} = \left(\widehat{\theta}_{i\bullet} - \widehat{\theta}_{i'\bullet}\right) \pm t_{\alpha/2;ddf_H}\sqrt{\frac{2}{JK}\left(MS_Y\left(TR\right) + \max\left(MS_Y\left(TC\right) - MS_Y\left(\varepsilon\right),0\right)\right)} \quad (21)$$

The analysis described so far treats both readers and cases as random factors, so it is termed *random-reader random-case* (RRRC). Special cases of the analysis, which regards either readers or cases as fixed factors, is possible, and the results are given in Appendix A.2. These are sometimes necessary if the number of readers or the number of cases is not large enough to support treating them as random factors (for example, one could have a single reader interpret a set of cases in two modalities and it would not make much sense to attempt to generalize this study to the population of readers).

## 2.5. ORH significance testing method

The statistical model underlying the OR method is (Obuchowski and Rockette 1995):

$$\begin{aligned} &\widehat{\theta}_{ij\{c\}} = \theta_0 + \Delta\theta_i + R_j + (\tau R)_{ij} + \varepsilon_{ij\{c\}} \\ &\sum_{i=1}^{I}\Delta\theta_i = 0 \end{aligned} \quad (22)$$

The left hand side is the estimated figure-of-merit $\widehat{\theta}_{ij\{c\}}$ for modality $i$ and *case-set* index $\{c\}$, where $c = 1, 2, \ldots, C$ denote different case sets (i.e., different *collections* of cases, not individual cases, emphasized by the curly bracket notation) sampled from the patient population). In practice the dataset is limited to $c = 1$, but resampling and other methods, are available to estimate the case-sample variability from a single case set realization. The first two terms on the right hand side of Eqn. 22 have their usual meanings. The remaining terms are mutually independent random samples: $R_j$ denotes a random contribution to the figure-of-merit of reader $j$, modeled as a sample from a zero-mean normal distribution with variance $\sigma_R^2$; $(\tau R)_{ij}$ denotes a treatment-dependent random contribution of reader $j$ in modality $i$, modeled as a sample from a zero-mean normal distribution with variance $\sigma_{\tau R}^2$. [We are abusing the notation but it is implicit that the variances in the OR model refer to the FOM, while those in the DBM model apply to pseudovalues.] The error term is modeled by a zero mean vector multivariate normal distribution with covariance matrix $\Sigma$ described by 4 parameters, $Var, Cov_1, Cov_2, Cov_3$, defined as follows:

$$Cov\left(\varepsilon_{ij\{c\}}, \varepsilon_{i'j'\{c\}}\right) = \left\{\begin{array}{ll} Var & i = i', j = j' \\ Cov_1 & i \neq i', j = j' \\ Cov_2 & i = i', j \neq j' \\ Cov_3 & i \neq i', j \neq j' \end{array}\right. \quad (23)$$

OR have suggested that the 4 elements of the covariance matrix should be ordered as follows:

$$Var \geq Cov_1 \geq Cov_2 \geq Cov_3 \quad (24)$$

Resampling methods are used to estimate the parameters of the covariance matrix. Using the bootstrap method (Efron and Tibshirani 1993), where $\{b\}$ is the $b^{th}$ bootstrap replicate,

$b = 1, 2, \ldots, B$,

$$\widehat{Cov}\left(\varepsilon_{ij\{c\}}, \varepsilon_{i'j'\{c\}}\right) = \left\langle \frac{1}{B-1} \sum_{b=1}^{B} \left(\theta_{ij\{b\}} - \theta_{ij\{\bullet\}}\right)\left(\theta_{i'j'\{b\}} - \theta_{i'j'\{\bullet\}}\right) \right\rangle_{ij} \tag{25}$$

As with the case-set index $\{c\}$, the bootstrap index $\{b\}$ denotes a set of cases. The averages, indicated by the bracket symbols, over modalities and readers are necessary since the covariances in the OR model are assumed to be independent of modality and reader. The jackknife estimate is:

$$\widehat{Cov}\left(\varepsilon_{ij\{c\}}, \varepsilon_{i'j'\{c\}}\right) = \left\langle \frac{K-1}{K} \sum_{k=1}^{K} \left(\theta_{ij(k)} - \theta_{ij(\bullet)}\right)\left(\theta_{i'j'(k)} - \theta_{i'j'(\bullet)}\right) \right\rangle_{ij} \tag{26}$$

DeLong, DeLong, and Clarke-Pearson (1988) have described an analytical covariance estimation method that is applicable as long as one restricts to the ROC paradigm and the Wilcoxon FOM (the bootstrap and the jackknife are applicable to any FOM).

Because of the correlated structure of the error term a customized ANOVA is needed. The null hypothesis is that the true figure-of-merit of all modalities are identical, i.e.,

$$NH : \Delta\theta_i = 0 \, (i = 1, 2, ..., I) \tag{27}$$

A modified $f$ statistic is needed, denoted $F^*_{ORH}$ and defined by (this is different from that originally suggested by OR):

$$F^*_{ORH} = \frac{MS\,(T)}{MS\,(TR) + H\left(J\left(\widehat{Cov}_2 - \widehat{Cov}_3\right)\right)} \tag{28}$$

Eqn. 28 incorporates Hillis' modification, which ensures that the constraint $Cov_2 \geq Cov_3$ is always obeyed. The mean square $(MS)$ terms are defined by (note the lack of the Y subscript, as these are calculated directly using FOM values):

$$\begin{aligned} MS\,(T) &= \tfrac{J}{I-1} \sum_{i=1}^{I} \left(\widehat{\theta}_{i\bullet} - \widehat{\theta}_{\bullet\bullet}\right)^2 \\ MS\,(TR) &= \tfrac{1}{(I-1)(J-1)} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\widehat{\theta}_{ij} - \widehat{\theta}_{i\bullet} - \widehat{\theta}_{\bullet j} + \widehat{\theta}_{\bullet\bullet}\right)^2 \end{aligned} \tag{29}$$

The observed statistic $F^*_{ORH}$ is distributed as an $f$ statistic with $ndf = I - 1$ and $ddf_H^{OR}$ degrees of freedom:

$$F^*_{ORH} \sim F_{ndf, ddf_H^{OR}} \tag{30}$$

where

$$ddf_H^{OR} = \frac{\left[MS\,(TR) + \left(J\left(\widehat{Cov}_2 - \widehat{Cov}_3\right)\right)\right]^2}{\frac{[MS(TR)]}{(I-1)(J-1)}^2} \tag{31}$$

As long as the jackknife is used to estimate the variance-components and co-variances in the two models, respectively, and the Wilcoxon or Wilcoxon-like statistic, the definitions of $ddf_H$ and $ddf_H^{OR}$ (Eqn. 19 and Eqn. 31) are the same; if the DeLong method or the bootstrap are used to estimate the co-variances, the two will yield slightly different results. The critical value of the $f$ statistic for rejection of the null hypothesis is given by $F_{1-\alpha,ndf,ddf_H^{OR}}$. The $p$ value of the test is given by:

$$p = P\left(F > F_{ORH}^{*} | F \sim F_{ndf,ddf_H^{OR}}\right) \tag{32}$$

The $(1-\alpha)\,100$ percent confidence interval for $\left(\widehat{\theta}_i - \widehat{\theta}_{i'}\right)$ is given by

$$CI_{1-\alpha} = \left(\widehat{\theta_{i\bullet}} - \widehat{\theta_{i'\bullet}}\right) \pm t_{\alpha/2;ddf_H^{OR}}\sqrt{\frac{2}{J}\left(MS\left(TR\right) + J\max\left(Cov_2 - Cov_3, 0\right)\right)} \tag{33}$$

The analysis described so far treats both readers and cases as random factors (RRRC). Special cases of the analysis, which regards either readers or cases as fixed factors, are given in Appendix A.3.

## 2.6. Sample size estimation for ROC studies

We will illustrate the procedure for the ORH method. Two modalities are assumed. The *observed effect size* (absolute value of the difference in FOMs between the two modalities) is $2\left|\widehat{\tau}_1\right|$. Under the alternative hypothesis $AH : \tau_i \neq 0$ the test statistic is distributed as a *non-central* F-distribution with $ndf = 1$ and to-be-determined $ddf$ and non-centrality parameter $\Delta$. The following sample size procedure (Hillis, Obuchowski, and Berbaum 2011) assumes random-readers and random cases; different formulae apply when either readers or cases is treated as a fixed effect, see below.

1. Specify the effect size $d$: typically, when dealing with area under the ROC curve as the FOM, one might choose observed effect size in the pilot study .

2. Estimate the OR modality-reader interaction variance component: this is given by (see Table 1 in Hillis (2007)):

   $$\hat{\sigma}_{\tau R}^2 = MS\left(TR\right) - \widehat{Var} + \widehat{Cov_1} + H\left(\widehat{Cov_2} - \widehat{Cov_3}\right)$$

   If this yields a negative variance, Hillis suggests setting it to zero.

3. Estimate the non-centrality parameter and the $ddf$ of the F-distribution. Let $K^*$ denote the number of cases in the pilot study, and let $J, K$ be the numbers of readers, cases in the pivotal study. The non-centrality parameter $\Delta$ and the $ddf$ are estimated by:

   $$\widehat{\Delta} = \frac{J\frac{d^2}{2}}{\hat{\sigma}_{\tau R}^2 + \frac{K^*}{K}\left(\widehat{Var} - \widehat{Cov_1} + \left(J-1\right)H\left(\widehat{Cov_2} - \widehat{Cov_3}\right)\right)}$$

   $$\widehat{ddf} = \left(J-1\right)\frac{\left[\hat{\sigma}_{\tau R}^2 + \frac{K^*}{K}\left(\widehat{Var} - \widehat{Cov_1} + \left(J-1\right)H\left(\widehat{Cov_2} - \widehat{Cov_3}\right)\right)\right]^2}{\left[\widehat{\sigma_{\tau R}^2} + \frac{K^*}{K}\left(\widehat{Var} - \widehat{Cov_1} + \left(J-1\right)H\left(\widehat{Cov_2} - \widehat{Cov_3}\right)\right)\right]^2}$$

4. The statistical power $1 - \beta$ at significance level $\alpha$ can be calculated using:

$$1 - \beta = P\left(F > F_{1-\alpha;1,\widehat{ddf}} | F \sim F_{1,\widehat{ddf};\widehat{\Delta}}\right)$$

   $F_{1,ddf;\Delta}$ denotes the non-central F-distribution with degrees of freedom 1, $ddf$, and non-centrality parameter $\Delta$ and $F_{1-\alpha;1,ddf}$ is the critical value of F such that fraction of the $1 - \alpha$ central F distribution with degrees of freedom 1, $ddf$ is below the critical value.

5. If the power is below the desired or target power, typically chosen to be 0.8, one tries successively larger value of $K$ until the target power is reached.

6. The procedure is repeated with different values of $J$ (depending on cost and other practicality issues, it might be better to have more reader each reading fewer cases to achieve the same target power).

Hillis has also described a procedure, currently unimplemented in **RJafroc**, for correcting the estimate if the numbers of non-diseased to diseased case ratio is substantially different between pilot and pivotal studies (Hillis *et al.* 2011).

*Formulae for fixed reader random case (FRRC) sample size estimation*

The only change needed is to define:

$$ddf = K - 1 \tag{34}$$

*Formulae for random reader fixed case (RRFC) sample size estimation*

The only change needed is to define:

$$ddf = J - 1 \tag{35}$$

# 3. Examples

It is assumed that the package has been installed from the CRAN website and loaded using the `library()` function. Users familiar with the Windows JAFROC graphical user interface (GUI), who simply wish to analyze their data, can skip to subsection 3.11, which describes the GUI.

## 3.1. Structure of the `dataset`

The package comes pre-loaded with three datasets: (1) an ROC dataset named `rocData`, which has been repeatedly used by Berbaum, Hillis and colleagues to illustrate advances in ROC methodology (Hillis 2007) (and referred to in their papers as Van Dyke data (Van Dyke, White, Obuchowski, Geisinger, Lorig, and Meziane 1993)), (2) an FROC dataset named `frocData`, contributed by Dr. Zanca (Zanca, Jacobs, Van Ongeval, Claus, Celis, Geniets, Provost, Pauwels, Marchal, and Bosmans 2009), and a simulated ROI dataset named `roiData` (see Appendix A.4 for details regarding the ROI simulator). The dataset structures are shown below:

```
R> str(rocData)

List of 8
 $ NL          : num [1:2, 1:5, 1:114, 1] 1 3 2 3 2 2 1 2 3 2 ...
 $ LL          : num [1:2, 1:5, 1:45, 1] 5 5 5 5 5 5 5 5 5 5 ...
 $ lesionNum   : int [1:45] 1 1 1 1 1 1 1 1 1 1 ...
 $ lesionID    : num [1:45, 1] 1 1 1 1 1 1 1 1 1 1 ...
 $ lesionWeight: num [1:45, 1] 1 1 1 1 1 1 1 1 1 1 ...
 $ dataType    : chr "ROC"
 $ modalityID  : chr [1:2] "0" "1"
 $ readerID    : chr [1:5] "0" "1" "2" "3" ...

R> str(frocData)

List of 8
 $ NL          : num [1:2, 1:4, 1:200, 1:7] -Inf -Inf -Inf -Inf -Inf ...
 $ LL          : num [1:2, 1:4, 1:100, 1:3] 5 4 4 3 5 5 4 2 4 5 ...
 $ lesionNum   : int [1:100] 1 1 1 1 1 1 1 1 1 1 ...
 $ lesionID    : num [1:100, 1:3] 1 1 1 1 1 1 1 1 1 1 ...
 $ lesionWeight: num [1:100, 1:3] 1 1 1 1 1 1 1 1 1 1 ...
 $ dataType    : chr "FROC"
 $ modalityID  : chr [1:2] "4" "5"
 $ readerID    : chr [1:4] "1" "3" "4" "5"

R> str(roiData)

List of 8
 $ NL          : num [1:2, 1:5, 1:90, 1:4] 0.957 0.907 0.57 0.824 1.473 ...
 $ LL          : num [1:2, 1:5, 1:40, 1:4] 1.51 2.32 2.37 2.19 2.34 ...
 $ lesionNum   : int [1:40] 2 3 2 2 3 3 1 2 3 3 ...
 $ lesionID    : num [1:40, 1:4] 2 1 1 1 1 2 4 1 1 1 ...
 $ lesionWeight: num [1:40, 1:4] 0.5 0.333 0.5 0.5 0.333 ...
 $ dataType    : chr "ROI"
 $ modalityID  : chr [1:2] "1" "2"
 $ readerID    : chr [1:5] "1" "2" "3" "4" ...
```

The ROC dataset has two modalities, 5 readers, 69 non-diseased and 45 diseased cases. The FROC data set has two modalities, 4 readers, 100 non-diseased and 100 diseased cases. Since ROC and ROI data are special cases of FROC data, the same data structure is used to accommodate all of them. The `dataType` field can be `ROC`, `FROC` or `ROI`. For ROC data, for a given modality and reader, the FP ratings are in the first $K_1$ values of the third dimension of the NL array and the corresponding TP ratings are in the $K_2$ values of the third dimension of the LL array. The fourth dimension of the NL and LL arrays, only the first value of which is used for ROC ratings, corresponds to the location index $l_s s$, i.e., the multiple marks of a given type, NL ($s = 1$) or LL ($s = 2$), that are possible for FROC data. In the above example, the dimensions of the NL array shows that there is least one image in the dataset with 7 NL marks, while the dimensions of the LL array shows that there is at least one diseased image with 3 lesions. The lesionNum field is an array of length $K_2$ whose elements contain the number of lesions in the diseased cases, i.e., $N_{k_2 2}$. The lesionID field is an integer label (not necessarily consecutive or even positive) used to distinguish between different lesions on the same case. This is necessary when weighted FOMs are used, as it is necessary to keep track of which lesion is getting which rating in order to assign it the correct weight. For example, LL[1,1,1,2] is the rating assigned to the $2^{nd}$ lesion for the first diseased case, first reader in the first modality and the corresponding label is lesionID[1, 2]. The lesionWeight field, corresponding to $W_{k_2 l_2}$, has the same dimensions as lesionID. The string vectors ModalityID and readerID are of length $I, J$, that are used to identify the modalities and readers, respectively. The ROI dataset has two modalities, 5 readers, 50 non-diseased and 40 diseased images, each with 4 ROIs. On the diseased images, the number of actually diseased ROIs varies from 1 to 4. The simulator, coded in R, is describded in Appendix A.4 and is available from http://www.devchakraborty.com/RoiData/RoiSimulator.zip.

## 3.2. Creating dataset objects

By adhering to the structure described above one can manually (or using code) create a dataset object (this could be useful in running simulation studies). For a single dataset it is more convenient to enter the data into an Excel sheet (both .xlsx and .xls files are supported) following the JAFROC data file format detailed in the help page for the package RJafroc-package and summarized below. A JAFROC Excel file contains three worksheets:

1. A Truth worksheet, which contains a list of all cases in the dataset and the number of lesions, if any, on each case, and the weight of each lesion.
2. A TP or LL worksheet (use TP for ROC data and LL for all other paradigms), which contains the ratings of TPs or LLs.
3. A FP or NL worksheet (use FP for ROC data and NL for all other paradigms), which contains the ratings of FPs or NLs.

For FROC data, except for the Truth worksheet, where each case must occur at least once, the number of rows in the other worksheets is variable. For ROC data each case appears once in the Truth worksheet and it appears once in either the FP or TP worksheet.

The ReadDataFile() function reads the data in JAFROC format (the default). If format = "MRMC", it will read .csv, .txt or .lrc files (http://perception.radiology.uiowa.edu/). If format = "iMRMC" it will read .imrmc files (https://code.google.com/p/imrmc/). In each case it returns a dataset object. The MRMC and iMRMC formats apply to ROC data only while the JAFROC format applies to all paradigms. The following code reads differently formattted datasets supplied with the package. The first group of six statements locates the full pathnames of the files in the user's installation, and the next group of six statements reads each file, with appropriate use of the format option.

```
R> rocXlsx <- system.file("tests", "rocData.xlsx", package = "RJafroc")
R> rocLrc <- system.file("tests", "rocData.lrc", package = "RJafroc")
R> rocCsv <- system.file("tests", "rocData.csv", package = "RJafroc")
```

```
R> rocImrmc <- system.file("tests", "rocData.imrmc", package = "RJafroc")
R> frocXlsx <- system.file("tests", "frocData.xlsx", package = "RJafroc")
R> roiXlsx <- system.file("tests", "roiData.xlsx", package = "RJafroc")
R>
R> RocDataXlsx<- ReadDataFile(fileName = rocXlsx)
R> RocDataLrc<- ReadDataFile(fileName = rocLrc, format = "MRMC")
R> RocDataCsv<- ReadDataFile(fileName = rocCsv, format = "MRMC")
R> RocDataImrmc<- ReadDataFile(fileName = rocImrmc, format = "iMRMC")
R> FrocDataXlsx<- ReadDataFile(fileName = frocXlsx)
R> RoiDataXlsx<- ReadDataFile(fileName = roiXlsx)
```

### 3.3. Analyzing an ROC dataset

One has two choices, DBMH significance testing, implemented by the function `DBMHAnalysis()`, or ORH significance testing, implemented by the function `ORHAnalysis()`. Both of these take a dataset object as the first argument, and have options for changing the significance level $\alpha$ of the test (the default is 0.05), and which factors (readers and/or cases) are to be regard as random (the default is `ALL`). Both functions use the weighted JAFROC FOM as the default, so to analyze ROC and ROI paradigm data one must explicitly specify the FOM options as shown below. The code below demonstrates the application of the DBMH significance testing procedure to an ROC `dataset` object:

```
R> retDbmRoc  <- DBMHAnalysis(rocData, fom = "Wilcoxon")
R> print(retDbmRoc)
```

The returned object, a list of 22 elements, has the structure shown in Table 4. The output can be understood if one uses the following abbreviations, often used in combination: `Trt` = treatment, `Rdr` = reader, `RRRC` = random reader random case, `FRRC` = fixed reader random case, `RRFC` = random reader fixed case, `ci` = $1-\alpha$ confidence interval, `fomArray` = $\widehat{\theta}_{ij}$, `f` = value of observed $f$ statistic, `p` = $p$ value for rejecting the null hypothesis, `DiffTrt` = reader-averaged FOM differences between pairs of modalities, `ddf` = denominator degrees of freedom for F-test (the numerator degrees of freedom is always $I-1$), `AvgRdrEachTrt` = the FOM is averaged over all readers, separately for each treatment, `varComp` = the DBM pseudovalue variance components defined in connection with Eqn. 12. For the dataset shown, the reader-averaged difference between the two modalities is not significant for `RRRC` ($p = 0.0517$), but is significant if either reader ($p = 0.021$) or case ($p = 0.042$) is regarded as a fixed factor.

To perform ORH significance testing one uses the function `ORHAnalysis()`, which takes the same arguments as `DBMHAnalysis()`, and additional optional arguments allowing choice of the covariance estimation method: `CovEstMethod` = `Jackknife`, `Bootstrap` or `DeLong` (`Jackknife` is the default) and if the bootstrap method is selected one can optionally specify the number of bootstraps (default = 200). The function will generate an error if the DeLong method is selected with a FOM that is not ”Wilcoxon” statistic on an ROC dataset. The return value of the `ORHAnalysis()`is a list of 21 elements, Table 5, similar to that of `DBMHAnalysis()`, but instead of 6 pseudovalue derived variance components, it returns the elements of the covariance matrix ($Var, Cov1, Cov2, Cov3$) and the mean-squares and variance components for the reader and treatment-reader effects. The following example illustrates how the returned varainces, covariances and mean squares can be extracted.

| Variable name | Description |
|---|---|
| `fomArray` | The FOM array of each reader and modality. |
| `anovaY` | The ANOVA table of the pseudovalues. |
| `anovaYi` | The ANOVA table of the pseudovalues for each modality. |
| `varComp` | The table of DBM variance components estimates. |
| `fRRRC` | The $f$ statistic for testing the null hypothesis, for the RRRC condition. |
| `ddfRRRC` | The denominator degrees of freedom of the $f$ statistic, for the RRRC condition. |
| `pRRRC` | The $p$ value of the significance test of the NH for the RRRC condition. |
| `ciDiffTrtRRRC` | The confidence intervals and related tests for the reader-averaged FOM differences between pairs of modalities, for the RRRC condition. |
| `ciAvgRdrEachTrtRRRC` | The confidence intervals and related tests for reader averaged FOM in each modality, for the RRRC condition. |
| `fFRRC` | The $f$ statistic for testing the null hypothesis, for the FRRC condition. |
| `ddfFRRC` | The denominator degrees of freedom of the FRRC $f$ statistic. |
| `pFRRC` | The $p$ value of the significance test of the NH, for the FRRC condition. |
| `ciDiffTrtFRRC` | The confidence intervals and related tests for the reader-averaged FOM differences between pairs of modalities, for the FRRC condition. |
| `ciAvgRdrEachTrtFRRC` | The confidence intervals and related tests for reader averaged FOM in each modality, for the FRRC condition. |
| `ssAnovaEachRdr` | The sum of squares table of the ANOVA of the pseudovalues for each reader (based on the data only for the specified reader). |
| `msAnovaEachRdr` | The mean squares table of the ANOVA of the pseudovalues for each reader (based on the data only for the specified reader). |
| `ciDiffTrtEachRdr` | The confidence intervals and related tests of the FOM differences between pairs of modalities for each reader. |
| `fRRFC` | The $f$ statistic for testing the null hypothesis, for the RRFC condition. |
| `ddfRRFC` | The denominator degrees of freedom of the $f$ statistic, for the RRFC condition. |
| `pRRFC` | The $p$ value of the significance test of the NH, for the RRFC condition. |
| `ciDiffTrtRRFC` | The confidence intervals and related tests for the FOM differences between pairs of modalities, for the RRFC condition. |
| `ciAvgRdrEachTrtRRFC` | The confidence intervals and related tests for reader averaged FOM in each modality, for the RRFC condition. |

Table 4: The structure of the object `retDbmRoc` returned by `DBMHAnalysis`. See text for abbreviations.

```
R> retORRoc  <- ORHAnalysis(rocData, fom = "Wilcoxon")
R> print(retORRoc)
R> CovOR <- retORRoc$varComp
R> cov1 <- CovOR$varCov[3]
R> cov2 <- CovOR$varCov[4]
R> cov3 <- CovOR$varCov[5]
R> varEps <- CovOR$varCov[6]
R> msTR <- retORRoc$msTR
R> msT <- retORRoc$msT
```

```
R> print(CovOR)
```

```
                 varCov
Var(R)     0.0015349993
Var(T*R)   0.0002004025
COV1       0.0003466137
COV2       0.0003440748
COV3       0.0002390284
Var(Error) 0.0008022883
```

Table 6 summarizes the results of DBMH and ORH analysis, for the latter the results of using different covariance estimation methods are shown, and compared to results yielded by OR-DBM MRMC (the University of Iowa software). Since ORH yields similar results to DBMH (they are identical for the Wilcoxon FOM) henceforth we will only show results for DBMH.

## 3.4. Sample size estimation for ROC studies

For the ROC dataset analyzed above, since random reader random case analysis was unable to reject the null hypothesis, a sample size estimate may be of interest for the purpose of planning a future study. We equate the effect size to the magnitude of the observed effect size, 0.0438, which is our best information about the magnitude of the true effect size (if the modalities will be further optimized prior to the pivotal study, it may be reasonable to posit 0.05 as the true effect size, but choosing an unrealistic effect size is not advisable). The following commands perform DBH analysis and extracts the relevant pseudovalue variance components and effect size for sample size estimation (note since the default FOM is `wJAFROC`, and one wishes to analyze ROC data, explicit specification of the fom option is necessary):

```
R> retDbm  <- DBMHAnalysis(rocData, fom = "Wilcoxon")
R> effectSize <- retDbm$ciDiffTrtRRRC$Estimate
R> varYTR <- retDbm$varComp$varComp[3]
R> varYTC <- retDbm$varComp$varComp[4]
R> varYEps <- retDbm$varComp$varComp[6]
```

The function `SampleSizeGivenJ()` can be used to determine the number of cases necessary to achieve a specified target power (default 0.8) for different values of $J$. Since the pilot study was conducted with 5 readers and barely reached significance, it is of interest to try different values 6:10 as in the code snippet below:

| Variable name | Description |
|---|---|
| `fomArray` | The FOM array of each reader and modality. |
| `msT` | The treatment mean square. |
| `msTR` | The treatment-reader mean square. |
| `varComp` | The first two elements contain the reader and modality-reader variance components, the rest contain, in order, Cov1, Cov2, Cov3 and Var. |
| `fRRRC` | The $f$ statistic for testing the null hypothesis, for the RRRC condition. |
| `ddfRRRC` | The denominator degrees of freedom of the $f$ statistic, for the RRRC condition. |
| `pRRRC` | The $p$ value of the significance test of the NH for the RRRC condition. |
| `ciDiffTrtRRRC` | The confidence intervals and related tests for the reader-averaged FOM differences between pairs of modalities, for the RRRC condition. |
| `ciAvgRdrEachTrtRRRC` | The confidence intervals and related tests for reader averaged FOM in each modality, for the RRRC condition. |
| `fFRRC` | The $f$ statistic for testing the null hypothesis, for the FRRC condition. |
| `ddfFRRC` | The denominator degrees of freedom of the FRRC $f$ statistic. |
| `pFRRC` | The $p$ value of the significance test of the NH, for the FRRC condition. |
| `ciDiffTrtFRRC` | The confidence intervals and related tests for the reader-averaged FOM differences between pairs of modalities, for the FRRC condition. |
| `ciAvgRdrEachTrtFRRC` | The confidence intervals and related tests for reader averaged FOM in each modality, for the FRRC condition. |
| `varCovEachRdr` | Obuchowski-Rockette Variance and Cov1 estimates for each reader. |
| `ciDiffTrtEachRdr` | The confidence intervals and related tests of the FOM differences between pairs of modalities for each reader. |
| `fRRFC` | The $f$ statistic for testing the null hypothesis, for the RRFC condition. |
| `ddfRRFC` | The denominator degrees of freedom of the $f$ statistic, for the RRFC condition. |
| `pRRFC` | The $p$ value of the significance test of the NH, for the RRFC condition. |
| `ciDiffTrtRRFC` | The confidence intervals and related tests for the FOM differences between pairs of modalities, for the RRFC condition. |
| `ciAvgRdrEachTrtRRFC` | The confidence intervals and related tests for reader averaged FOM in each modality, for the RRFC condition. |

Table 5: The structure of the object `retORHRoc`returned by `ORHAnalysis`. See text for abbreviations.

| | Statistic | **RJafroc** | OR-DBM MRMC |
|---|---|---|---|
| | $\widehat{\theta}_{1\bullet}, \widehat{\theta}_{2\bullet}$ | 0.897, 0.941 | 0.897, 0.941 |
| DBMH | $\widehat{\theta}_{1\bullet} - \widehat{\theta}_{2\bullet}$ | -0.0438 | -0.0438 |
| | $p$ value | 0.0517 | 0.0517 |
| | $f$ statistic | 4.46 | 4.46 |
| | ddf | 15.3 | 15.26 |
| | Confidence interval | (-0.088, 0.000359) | (-0.088,0.00036) |
| ORH Jackknife | $\widehat{\theta}_{1\bullet} - \widehat{\theta}_{2\bullet}$ | -0.0438 | -0.0438 |
| | $p$ value | 0.0517 | 0.0517 |
| | $f$ statistic | 4.46 | 4.46 |
| | ddf | 15.3 | 15.26 |
| | Confidence interval | (-0.088, 0.000359) | (-0.088,0.00036) |
| ORH Bootstrap<br>boots = 200 | $\widehat{\theta}_{1\bullet} - \widehat{\theta}_{2\bullet}$ | -0.0438 | -0.0438 |
| | $p$ value | 0.0501 | 0.0558 |
| | $f$ statistic | 4.56 | 4.21 |
| | ddf | 14.5 | 17.07 |
| | Confidence interval | (-0.0876, 0.0000164) | (-0.0888, 0.00121) |
| ORH DeLong | $\widehat{\theta}_{1\bullet} - \widehat{\theta}_{2\bullet}$ | -0.0438 | -0.0438 |
| | $p$ value | 0.0512 | 0.0512 |
| | $f$ statistic | 4.48 | 4.48 |
| | ddf | 15.1 | 15.07 |
| | Confidence interval | (-0.0879, 0.000267) | (-0.0879,0.00027) |

Table 6: Results of DBMH and ORH analysis (with different methods for estimating the covariance matrix) for `rocData` compared to that yielded by OR-DBM MRMC (the University of Iowa Windows software). Only results for random readers and random cases are shown.

```
R> for (J in 6:10) {
+   ret <- SampleSizeGivenJ(J, varYTR, varYTC, varYEps,
+     effectSize = effectSize)
+   message("# of readers = ", J, ", estimated # of cases = ", ret$K, "\n",
+     "predicted power = ", signif(ret$power, 4), "\n")
+ }

# of readers = 6, estimated # of cases = 251
predicted power = 0.8005

# of readers = 7, estimated # of cases = 211
predicted power = 0.8008

# of readers = 8, estimated # of cases = 188
predicted power = 0.8007

# of readers = 9, estimated # of cases = 173
predicted power = 0.8005

# of readers = 10, estimated # of cases = 163
predicted power = 0.8016
```

This type of information can be used to test the practicality of different study designs. The preceding analysis assumed `RRRC`; to get results assuming fixed readers, supply the option `randomOption = "CASES"`; to get results assuming fixed cases, supply the option `randomOption = "READERS"`.

Similar analysis can be conducted using the ORH method.

```
R> retOR  <- ORHAnalysis(rocData, fom = "Wilcoxon")
R> effectSize <- retDbm$ciDiffTrtRRRC$Estimate
R> CovOR <- retOR$varComp
R> cov1 <- CovOR$varCov[3]
R> cov2 <- CovOR$varCov[4]
R> cov3 <- CovOR$varCov[5]
R> varErrOR <- CovOR$varCov[6]
R> msTR <- retOR$msTR
R> KStar <- length(rocData$NL[1,1,,1])
R> for (J in 6:10) {
+   ret <- SampleSizeGivenJ(J, cov1 = cov1, cov2 = cov2, cov3 = cov3,
+     varEps = varErrOR, msTR = msTR, KStar = KStar, effectSize = effectSize)
+   message("# of readers = ", J, ", estimated # of cases = ", ret$K, "\n",
+     "predicted power = ", signif(ret$power, 4), "\n")
+ }

# of readers = 6, estimated # of cases = 251
predicted power = 0.8005
```

```
# of readers = 7, estimated # of cases = 211
predicted power = 0.8008

# of readers = 8, estimated # of cases = 188
predicted power = 0.8007

# of readers = 9, estimated # of cases = 173
predicted power = 0.8005

# of readers = 10, estimated # of cases = 163
predicted power = 0.8016
```

These are identical to those obtained with DBMH analysis.

## 3.5. Analyzing an FROC dataset

Analysis of location specific data (free-response or ROI) is not fundamentally different from that of ROC paradigm data. As long as the figure of merit is a scalar significance testing methods developed for ROC apply to the selected FOM. We illustrate analysis of this dataset using the function `DBMHAnalysis()`, noting that `wJAFROC` is the default FOM. There are 100 non-diseased and 100 diseased images in the pre-loaded FROC dataset, with the number of lesions on the diseased images ranging from 1 to 3. The two modalities are labeled 4 and 5 (the full dataset, containing data for 5 modalities, is available from author DPC). The following example illustrates application of `DBMHAnalysis()` to the `frocData dataset` object. Since the default FOM is `wJAFROC`, explicit specification of the `fom` option is unnecessary.

```
R> retDbmwJafroc  <- DBMHAnalysis(frocData)
R> print(retDbmwJafroc)
```

The following three examples illustrate other analyses possible with free-response data, with appropriate specification of the `fom` option. The first example analyzes the data using the `wJAFROC1` FOM, which uses the highest rated NL mark on diseased cases and gives equal importance to all diseased cases. The second example uses the `JAFROC` FOM, which tends to give extra importance to cases with many lesions but ignores NL marks on diseased cases. The third example illustrates usage of the `JAFROC1` FOM, which uses the highest rated NL mark on diseased cases and gives extra importance to cases with many lesions. *The* `wJAFROC1` *and* `JAFROC1` *FOMs should only be used if the dataset contains no non-diseased cases.*

```
R> retDbmwJafroc1  <- DBMHAnalysis(frocData, fom = "wJAFROC1")
R> print(retDbmwJafroc1)
R>
R> retDbmJafroc  <- DBMHAnalysis(frocData, fom = "JAFROC")
R> print(retDbmJafroc)
R>
R> retDbmJafroc1  <- DBMHAnalysis(frocData, fom = "JAFROC1")
R> print(retDbmJafroc1)
```

Table 7 shows results of DBMH-analysis, using location specific figures of merit (`JAFROC`, `wJAFROC`, `JAFROC1` and `wJAFROC1`), applied to the free-response dataset and compared to results obtained using the Windows version V 4.2.1 of JAFROC software. Only results for random readers and random cases are shown. The reason for making `wJAFROC` the default FOM is that the software is primarily designed to analyze free-response data and by doing weighted analysis each diseased case gets the same importance in the analysis, regardless of the number of lesions in it. With un-weighted analysis, selected by setting the FOM option to `JAFROC` or `JAFROC1`, the results can be skewed by cases having a large number of lesions (we have encountered a nuclear medicine bone-scan dataset where the number of lesions per patient varied from a few to a hundred).

The `JAFROC1` FOMs use all highest rated NL marks, even those on diseased cases. While `JAFROC1` may give higher statistical power, it mixes two types of discriminability, that between LLs and NLs on normal cases (clinically very important) and that between LLs and NLs on abnormal cases (clinically less important). For this reason we do not recommend `JAFROC1` or `wJAFROC1`, unless the dataset has no non-diseased cases, in which situation the mixing effect just referred to cannot occur. Another issue with the `JAFROC1` and `wJAFROC1` FOMs is that they will depend on the case mix (i.e., the proportion of cases that are actually diseased). This means that two investigators sampling the same population but using different case mixes may get different results, even after sampling effects are accounted for. This issue also applies to the ROI paradigm. For these reasons we prefer `JAFROC`, and particularly `wJAFROC` FOMs for characterizing free-response performance.

Inferred ROC analysis can be performed on free-response data. The following three examples are for ROC data inferred from FROC data using different methods of inferring the data: the first example uses the highest rating method, the second uses the SongA1 method (average rating) and the third example uses the SongA2 method (stochastic dominance).

```
R> retDbmHrAuc   <- DBMHAnalysis(frocData, fom = "HrAuc")
R> retDbmSongA1  <- DBMHAnalysis(frocData, fom = "SongA1")
R> retDbmSongA2  <- DBMHAnalysis(frocData, fom = "SongA2")
```

Table 8 shows results of DBMH-analysis, using inferred ROC figures of merit (`HrAuc`, `SongA1` and `SongA2`), applied to a free-response dataset and compared to results obtained using the Windows version of the software. Only results for random readers and random cases are shown. The Song FOMs, particularly A2, are computationally quite intensive (to put it in perspective, software run times in this field pale in comparison to the effort required to acquire the data, often 6 months or more).

Besides showing that the package gives identical results to JAFROC, the results illustrate some general principles. (1) While all methods reject the NH, the $p$ value is considerably smaller for weighted JAFROC (6.46e-06) as compared to the inferred ROC methods (range 0.0095 to 0.024). While one cannot infer statistical power from a comparison of $p$ values on a single dataset, the increased statistical power of JAFROC analysis has been confirmed with simulation studies (Chakraborty 2002; Chakraborty and Berbaum 2004; Chakraborty 2008) and is one reason this paradigm is gaining acceptance. (2) The `JAFROC` FOM for each modality is smaller than the corresponding inferred ROC FOMs. This is because of the localization requirement, which implies that LLF is always less than the corresponding inferred TPF. In other words lesions are only counted towards LLF if they are correctly localized, while TPF

| FOM | Statistic | **RJafroc** | JAFROC V4.2.1 |
|---|---|---|---|
| wJAFROC | $\widehat{\theta}_{4\bullet}, \widehat{\theta}_{5\bullet}$ | 0.768, 0.714 | 0.768, 0.714 |
| | $\widehat{\theta}_{4\bullet} - \widehat{\theta}_{5\bullet}$ | 0.0548 | 0.0548 |
| | $CI\left(\widehat{\theta}_{4\bullet} - \widehat{\theta}_{5\bullet}\right)$ | (0.0328, 0.0769) | (0.0328, 0.0769) |
| | $p$ value | 6.46E-06 | <0.0001 |
| | $f$ statistic | 24.9 | 24.88 |
| | ddf | 54.96 | 54.96 |
| JAFROC | $\widehat{\theta}_{4\bullet}, \widehat{\theta}_{5\bullet}$ | 0.758, 0.703 | 0.758, 0.703 |
| | $\widehat{\theta}_{4\bullet} - \widehat{\theta}_{5\bullet}$ | 0.0548 | 0.0548 |
| | $CI\left(\widehat{\theta}_{4\bullet} - \widehat{\theta}_{5\bullet}\right)$ | (0.0315, 0.0780) | (0.0316, 0.0780) |
| | $p$ value | 5.63E-06 | <0.0001 |
| | $f$ statistic | 21.6 | 21.6 |
| | ddf | 236.4 | 236.4 |
| wJAFROC1 | $\widehat{\theta}_{4\bullet}, \widehat{\theta}_{5\bullet}$ | 0.783, 0.729 | 0.783, 0.729 |
| | $\widehat{\theta}_{4\bullet} - \widehat{\theta}_{5\bullet}$ | 0.054 | 0.054 |
| | $CI\left(\widehat{\theta}_{4\bullet} - \widehat{\theta}_{5\bullet}\right)$ | (0.036, 0.0715) | (0.036, 0.0715) |
| | $p$ value | 1.91E-09 | <0.0001 |
| | $f$ statistic | 36.5 | 36.51 |
| | ddf | 1491 | 1492 |
| JAFROC1 | $\widehat{\theta}_{4\bullet}, \widehat{\theta}_{5\bullet}$ | (0.773, 0.720) | (0.773, 0.720) |
| | $\widehat{\theta}_{4\bullet} - \widehat{\theta}_{5\bullet}$ | 0.0535 | 0.0535 |
| | $CI\left(\widehat{\theta}_{4\bullet} - \widehat{\theta}_{5\bullet}\right)$ | (0.0291, 0.0779) | (0.0291, 0.078) |
| | $p$ value | 5.55E-05 | <0.0001 |
| | $f$ statistic | 19.3 | 19.4 |
| | ddf | 51.07 | 51.07 |

Table 7: Results of DBMH-analysis, using location specific figures of merit, applied to a free-response dataset and compared to results obtained using the Windows version of the software. Only results for random readers and random cases are shown.

| FOM | Statistic | **RJafroc** | JAFROC V4.2 |
|---|---|---|---|
| `HrAuc` | $\widehat{\theta}_{4\bullet}, \widehat{\theta}_{5\bullet}$ | 0.851, 0.808 | 0.851, 0.808 |
| | $\widehat{\theta}_{4\bullet} - \widehat{\theta}_{5\bullet}$ | 0.04219 | 0.04219 |
| | $CI\left(\widehat{\theta}_{4\bullet} - \widehat{\theta}_{5\bullet}\right)$ | (0.0098, 0.0746) | (0.0098, 0.0746) |
| | $p$ value | 0.0240 | 0.0240 |
| | $f$ statistic | 14.96 | 14.96 |
| | ddf | 3.429 | 3.43 |
| `SongA1` | $\widehat{\theta}_{4\bullet}, \widehat{\theta}_{5\bullet}$ | 0.853, 0.808 | 0.853, 0.808 |
| | $\widehat{\theta}_{4\bullet} - \widehat{\theta}_{5\bullet}$ | 0.04505 | 0.04505 |
| | $CI\left(\widehat{\theta}_{4\bullet} - \widehat{\theta}_{5\bullet}\right)$ | (0.0186, 0.0715) | (0.0186, 0.0715) |
| | $p$ value | 0.0095 | 0.0095 |
| | $f$ statistic | 23.1 | 23.1 |
| | ddf | 3.84 | 3.84 |
| `SongA2` | $\widehat{\theta}_{4\bullet}, \widehat{\theta}_{5\bullet}$ | 0.847, 0.800 | 0.847, 0.800 |
| | $\widehat{\theta}_{4\bullet} - \widehat{\theta}_{5\bullet}$ | 0.0468 | 0.0468 |
| | $CI\left(\widehat{\theta}_{4\bullet} - \widehat{\theta}_{5\bullet}\right)$ | (0.0156, 0.0780) | (0.0156, 0.0780) |
| | $p$ value | 0.0173 | 0.0173 |
| | $f$ statistic | 22.5 | 22.53 |
| | ddf | 3.03 | 3.03 |

Table 8: Results of DBMH-analysis, using inferred ROC FOMs (`HrAuc`, `SongA1` and `SongA2`), applied to the included free-response dataset

is only concerned with the inferred single rating per case. (3) The effect size is larger for JAFROC (0.0548) than for any of the inferred ROC methods (0.047 for Song A2). The reason for the larger JAFROC effect size is that the figure of merit has a larger range over which it can vary, 0 to 1, while any ROC FOM is restricted to the range 0.5 to 1. Since effect size appears as the square in sample size calculations, this contributes towards JAFROC's higher statistical power.

## 3.6. Analyzing an ROI dataset

The package comes pre-loaded with an ROI dataset, `roiData`. The `NL[1:2, 1:5, 1:90, 1:4]` array contains the ratings of all non-diseased ROIs while the `LL[1:2, 1:5, 1:90, 1:4]` array contains the ratings of all diseased ROIs. Since `wJAFROC` is the default FOM, one needs to explicitly specify the `ROI` FOM when using the function `DBMHAnalysis()` to analyze ROI data. The following example illustrates usage of `DBMHAnalysis()` to analyze an ROI `dataset` object. Note that explicit specification of the `fom` option is necessary.

```
R> retDbmRoi  <- DBMHAnalysis(roiData, fom = "ROI")
```

The results of RRRC analysis using **RJafroc** and C++ version of JAFROC are summarized in Table 9.

| FOM | Statistic | **RJafroc** | JAFROC V4.2 |
|---|---|---|---|
| `ROI` | $\widehat{\theta}_{4\bullet}, \widehat{\theta}_{5\bullet}$ | 0.884, 0.922 | 0.884, 0.922 |
| | $\widehat{\theta}_{4\bullet} - \widehat{\theta}_{5\bullet}$ | -0.038 | -0.038 |
| | $CI\left(\widehat{\theta}_{4\bullet} - \widehat{\theta}_{5\bullet}\right)$ | (-0.064, -0.0116) | (-0.064, -0.0116) |
| | $p$ value | 0.00823 | 0.00823 |
| | $f$ statistic | 9.687 | 9.69 |
| | ddf | 13.0 | 13.0 |

Table 9: DBMH applied to ROI data analysis. Only results for random readers and random cases are shown.

## 3.7. Generating an output report

The function `OutputReport()` is used to to analyze data and generate a formatted report closely patterned on that of OR-DBM MRMC and DBM-MRMC. The following example illustrates usage of this function for the included ROC dataset object `rocData`; the dataset is to be analyzed using the DBMH method and the Wilcoxon statistic as FOM. Since the default `method` is DBMH, it does not need explicit specification. However, the `fom` option needs to be explicitly specified, since the default is `wJAFROC`. The `showWarnings` option is set to `FALSE` as otherwise the program will pause for user input if the named output file already exists. The `dataDscrpt` option supplies a plain English description of the dataset. It is only needed if a `dataset` object is specified. The default is the variable name of the `dataset` object.

```
R> OutputReport(dataset = rocData, fom = "Wilcoxon",
+   dataDscrpt = "MyROCData", showWarnings = FALSE)
```

The next example explicitly specifies the name of the output report file, using the reportFile option. If this option is missing, the function will use the file name of the data file or the value of the dataDscrpt option followed by the underscore separated concatenation of method and fom as the output file name.

```
R> OutputReport(dataset = rocData, fom = "Wilcoxon",
+   reportFile = "MyROCDataAnalysis.txt",  showWarnings = FALSE)
```

The next example applies the ORH method to analyzing the data; since this is different from the default (DBMH) the option needs explicit specification.

```
R> OutputReport(dataset = rocData, method = "ORH", fom = "Wilcoxon",
+   showWarnings = FALSE)
```

The next example is deliberately included as an example of erroneous usage of the function. It will generate an error since the fom option set to Wilcoxon is incompatible with the FROC dataset (see Table 1 for valid FOMs with FROC data).

```
R> OutputReport(dataset = frocData, fom = "Wilcoxon",
+   showWarnings = FALSE)
```

The next example illustrates valid usage assuming wJAFROC as the FOM; the fom option does not need to be explicitly specified, since wJAFROC is the default.

```
R> OutputReport(dataset = frocData, method = "ORH",
+   showWarnings = FALSE)
```

The next example illustrates analysis of FROC using the trapezoidal area under the highest rating inferred ROC curve as the FOM.

```
R> OutputReport(dataset = frocData, fom = "HrAuc",
+   showWarnings = FALSE)
```

The final example illustrates analysis of the include ROI dataset using the ROI FOM, which needs to be explcitly specified using the fom option.

```
R> OutputReport(dataset = roiData, method = "ORH", fom = "ROI",
+   showWarnings = FALSE)
```

Alternatively, one can skip the dataset object creation step: the following example reads the data file, analyzes it and generates the output report.

```
R> OutputReport("rocData.xlsx", format = "JAFROC", method = "DBMH",
+   fom = "Wilcoxon", dataDscrpt = "MyROC2Data",  showWarnings = FALSE)
```

## 3.8. Saving a data file in a specified format

The function `SaveDataFile()` can be used to save an ROC dataset object in any compatible format, thereby allowing it to be analyzed with alternate software. The following examples illustrate its usage (the OR-DBM MRMC specified “*.csv” and “*.txt” files are identical except for the different file extensions).

```
R> SaveDataFile(dataset = rocData, fileName = "rocData2.xlsx",
+   format = "JAFROC")
R> SaveDataFile(dataset = rocData, fileName = "rocData2.csv", format = "MRMC")
R> SaveDataFile(dataset = rocData, fileName = "rocData2.lrc", format = "MRMC",
+   dataDscrpt = "ExampleROCdata")
R> SaveDataFile(dataset = rocData, fileName = "rocData2.txt", format = "MRMC",
+   dataDscrpt = "ExampleROCdata2")
R> SaveDataFile(dataset = rocData, fileName = "rocData.imrmc",
+   format = "iMRMC", dataDscrpt = "ExampleROCdata3")
```

## 3.9. ROC data visualization

The package includes a function `EmpiricalOpCharac()` for plotting trapezoidal ROC curves. The following commands will create trapezoidal ROC curves for all combinations of modalities and readers in the `rocData` dataset:

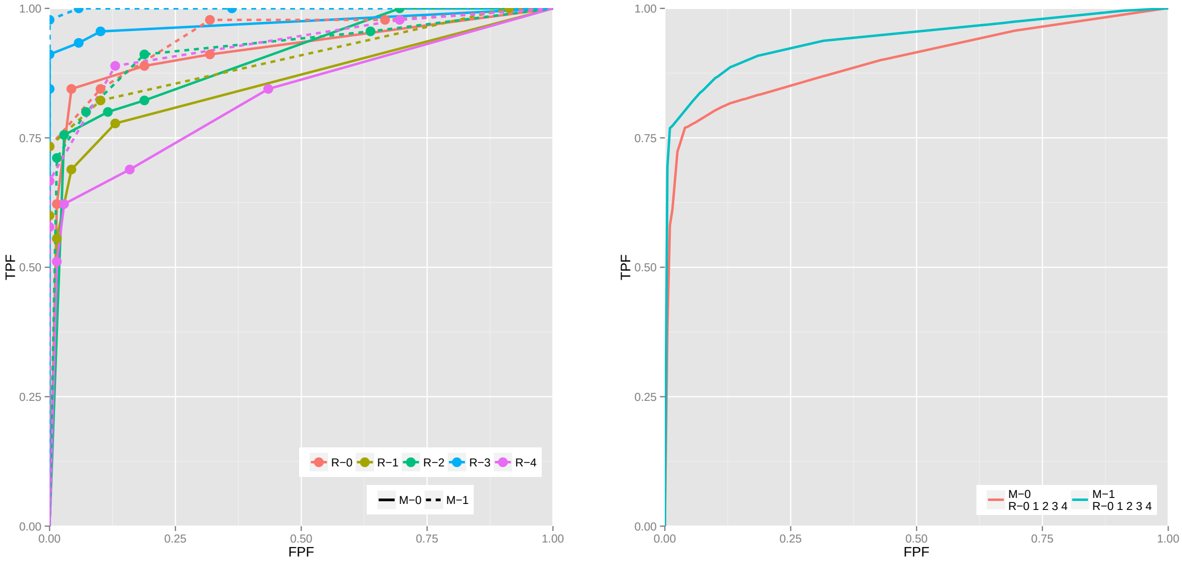
```
R> plotM <- c(1:2)
R> plotR <- c(1:5)
R> plotROC <- EmpiricalOpCharac(data = rocData, trts = plotM, rdrs = plotR,
+   opChType = "ROC")
```

The `trts = plotM` argument tells the function to plot both modalities, and `rdrs = plotR` tells it to plot data for all five reader in each modality. The result of printing `plotROC`, a **ggplot2** object (Wickham 2009), is Fig. 1. Since ROC analysis is a subspecialty of statistics, and not all users may be familiar with it, we point out the obvious: an operating characteristic that approaches the top-left corner has greater area under the trapezoidal curve, which implies greater performance. The ROC curve for a guessing observer would be the diagonal line connecting $(0,0)$ to $(1,1)$.

Fig. 1(a) shows the large variability in performance between the readers, which is one reason one needs to adequately sample the reader population. The following construct can be used to plot operating characteristics for each modality, averaged over readers (Fig. 1(b)).

```
R> plotMAvg <- list(1, 2)
R> plotRAvg <- list(c(1:5),c(1:5))
R> plotRocAvg <- EmpiricalOpCharac(dataset = rocData, trts = plotMAvg,
+   rdrs = plotRAvg, opChType = "ROC")
```

This tells the function to create two plots, one per modality, where each plot is averaged over all 5 readers.



(a)

(b)

Figure 1: (a) shows empirical receiver operating characteristics for all 5 readers in both modalities. (b) shows reader-averaged receiver operating characteristics for the two modalities.

## 3.10. Free-response data visualization

The function `EmpiricalOpCharac()` can be used to plot trapezoidal ROC/AFROC/FROC curves. The following commands will create trapezoidal ROC curves for all 8 combinations of 2 modalities and 4 readers in the `frocData` dataset, Fig. 2(a), and reader-averaged ROC, Fig. 2(b), reader-averaged AFROC, Fig. 2(c) and reader-averaged FROC curves, Fig. 2(d).

```
R> plotM <- c(1:2)
R> plotR <- c(1:4)
R> plotROC <- EmpiricalOpCharac(data = frocData, trts = plotM, rdrs = plotR,
+   opChType = "ROC")
R>
R> plotMAvg <- list(1, 2)
R> plotRAvg <- list(c(1:4),c(1:4))
R> plotRocAvg <- EmpiricalOpCharac(data = frocData, trts = plotMAvg,
+   rdrs = plotRAvg, opChType = "ROC")
R>
R> plotMAvg <- list(1, 2)
R> plotRAvg <- list(c(1:4),c(1:4))
R> plotAFROC <- EmpiricalOpCharac(data = frocData, trts = plotMAvg,
+   rdrs = plotRAvg, opChType = "AFROC")
R>
R> plotMAvg <- list(1, 2)
R> plotRAvg <- list(c(1:4),c(1:4))
R> plotFROC <- EmpiricalOpCharac(data = frocData, trts = plotMAvg,
+   rdrs = plotRAvg, opChType = "FROC")
```

Panel (a) does show, for each reader, coded by color, that the dotted lines are above the corresponding solid lines. This is confirmed in the averaged ROC, AFROC and FROC curves (panels (b), (c) and (d)). Panel (c) shows the difference that was found to be significant by DBMH/ORH analysis using both wJAFROC and HrAuc figures of merit.

The numbering of the readers is not sequential; the reader IDs are actually string labels, and in this dataset for some reason the experimenter chose not to use the sequential labels 1 - 4. Comparing panels (b) and (c) one can appreciate that the AFROC curve is below the corresponding ROC curve, and that the difference is areas is larger for the AFROC than the ROC. Panel (d) shows the averaged FROC curves; although used by some investigators, this is a poor summary of performance. Even the partial area under the FROC to the left of some defined abscissa value is not a good FOM (Youden 1950; Hillis 2007), as it does not give credit for non-diseased images with no marks (these are actually high confidence correct decisions - i.e., perfect decisions).

## 3.11. Graphical user interface

It is assumed that **RJafroc** has been installed from the CRAN website and loaded using the library() function. The graphical user interface (GUI) is invoked by the function RJafrocGui():

```
R> RJafrocGui()
```

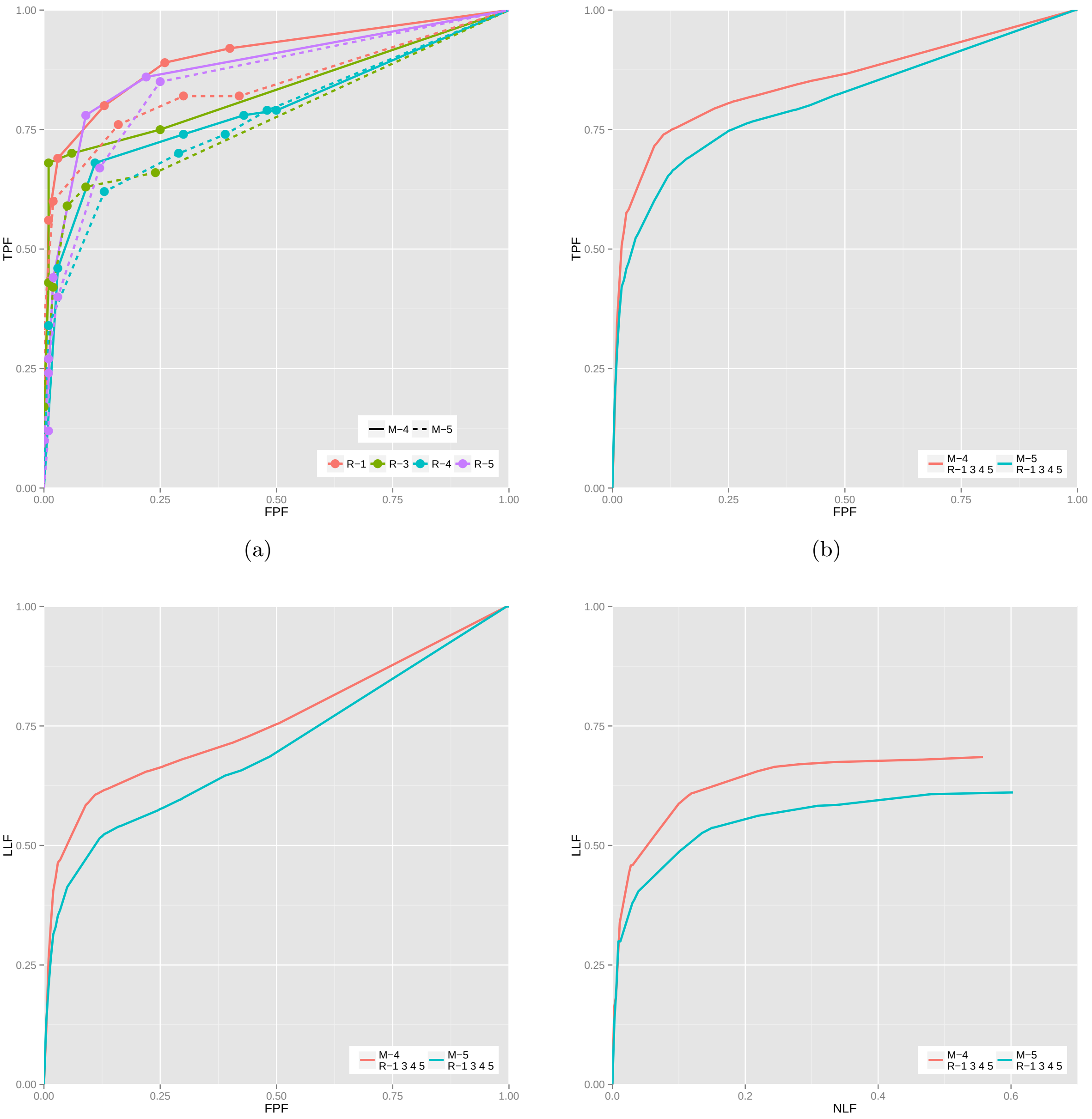
Due to a bug in RStudio's internal browser (Windows Version 0.99.467), which prevents saving plot files, Windows users may wish to invoke it as follows:

```
R> RJafrocGui(useBrowser = TRUE)
```

Fig. 3 shows the opening screen of the GUI. Clicking on "Choose File" opens a standard file-select window (Finder or Explorer) that allows one to select the desired data file. In Fig. 4, the included FROC data file "frocData.xlsx" has been selected (see subsection 3.1 for details of this dataset). Clicking on "Open" shows Fig. 5. The "Data Viewer" panel allows one to look at the data (this capability is not available in the Windows JAFROC software version 4.2.1). In Fig. 5 the "Truth" panel is shown. Clicking on NL or LL will show the corresponding ratings data for NLs and LLs, respectively. The "Data Conversion" panel allows the dataset to be saved in other formats, e.g., highest rating inferred ROC. Clicking on "Analysis" shows Fig. 6. Fig. 7 shows the result of clicking on the "Analysis" tab in Fig. 6. This screen shows a summary of the dataset, and allows selection of analysis method, the FOM and the significance level of the testing. Clicking on "Analyze" shows the results of the analysis, Fig. 7. The format closely follows that of Windows JAFROC. Clicking on the "Save Report" button allows the results to be saved to a text file.

## 3.12. Software comparison

Table 10 compares the features and capabilities of existing online software and **RJafroc**: data file format, whether they are open-source, and if so, the programming language expertise needed to understand them, whether they are cross platform applications, whether individual



(a)

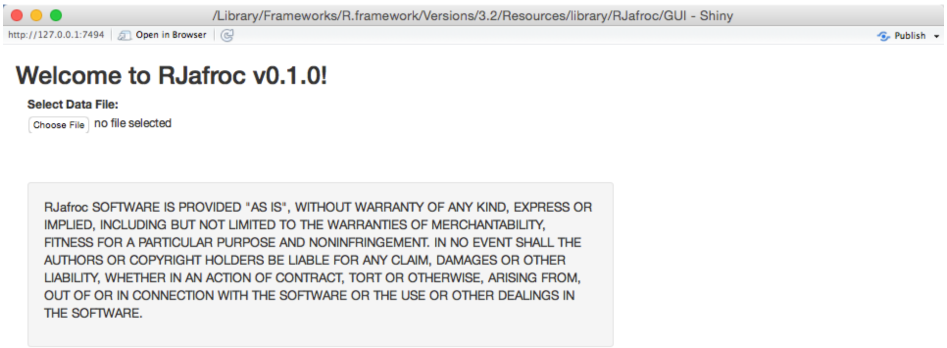
(b)

(c)

(d)

Figure 2: (a) shows the empirical highest rating inferred ROC curves for all combinations of modalities and readers. (b) shows the reader-averaged inferred ROC curves for both modalities. (c) shows the reader-averaged AFROC curves for both modalities. (d) shows the reader-averaged FROC curves for both modalities.

modules can be called from other languages, whether they include integrated visualization routines, and the degree to which they accommodate location paradigms.


/Library/Frameworks/R.framework/Versions/3.2/Resources/library/RJafroc/GUI - Shiny
http://127.0.0.1:7494
Open in Browser
Publish
Choose File

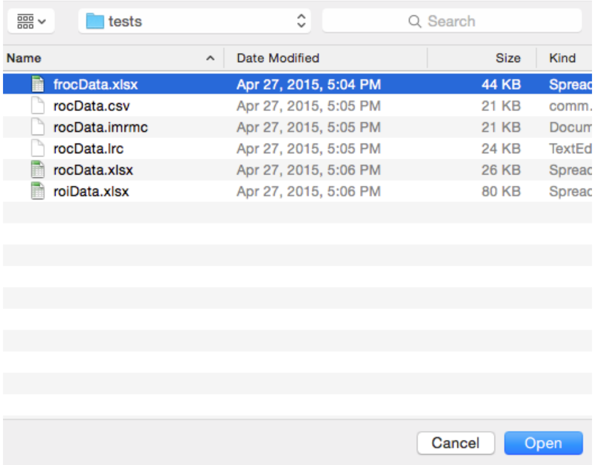
Welcome to RJafroc v0.1.0!

Select Data File:

no file selected

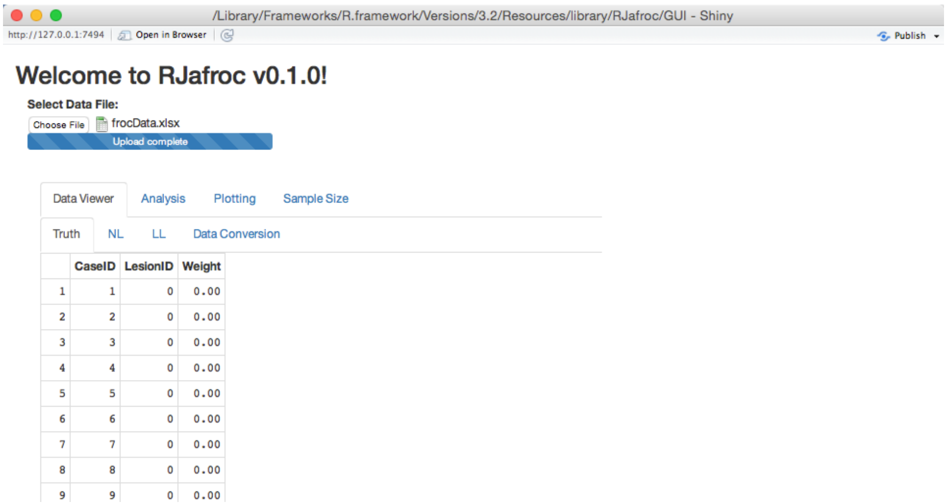
RJafroc SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Figure 3: The opening screen of the graphical user interface to **RJafroc**. The "Choose File" button allows selection of the data file to be analyzed.



| Name | Date Modified | Size | Kind |
|---|---|---|---|
| frocData.xlsx | Apr 27, 2015, 5:04 PM | 44 KB | Spread |
| rocData.csv | Apr 27, 2015, 5:05 PM | 21 KB | comm. |
| rocData.imrmc | Apr 27, 2015, 5:05 PM | 21 KB | Docum |
| rocData.lrc | Apr 27, 2015, 5:05 PM | 24 KB | TextEd |
| rocData.xlsx | Apr 27, 2015, 5:06 PM | 26 KB | Spread |
| roiData.xlsx | Apr 27, 2015, 5:06 PM | 80 KB | Spread |

Figure 4: Selecting the data file to be analyzed. All shown file formats are supported.



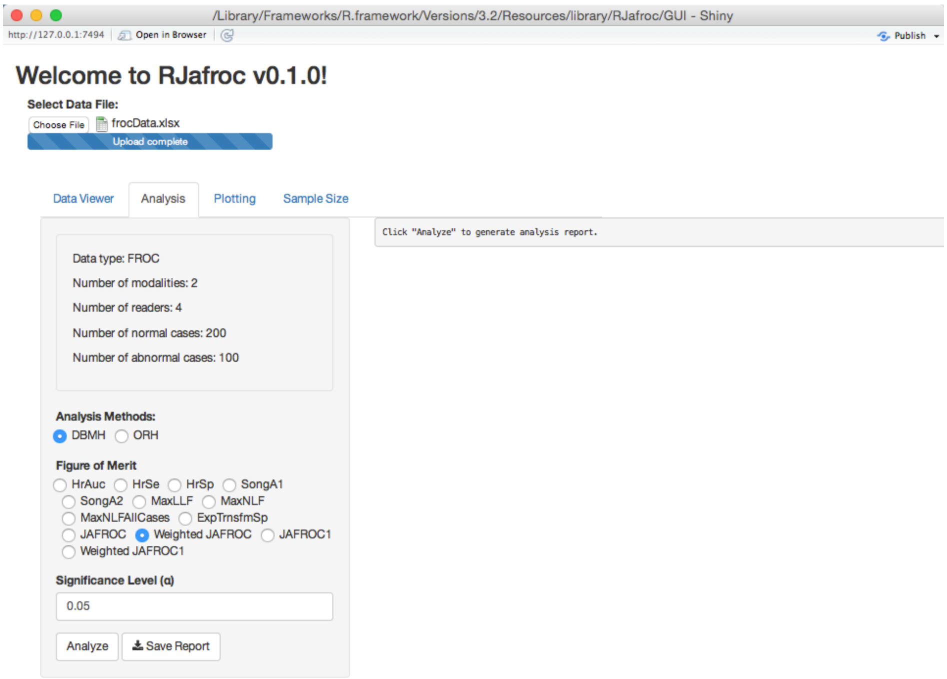
Welcome to RJafroc v0.1.0!

Select Data File:

frocData.xlsx

| | CaseID | LesionID | Weight |
|---|---|---|---|
| 1 | 1 | 0 | 0.00 |
| 2 | 2 | 0 | 0.00 |
| 3 | 3 | 0 | 0.00 |
| 4 | 4 | 0 | 0.00 |
| 5 | 5 | 0 | 0.00 |
| 6 | 6 | 0 | 0.00 |
| 7 | 7 | 0 | 0.00 |
| 8 | 8 | 0 | 0.00 |
| 9 | 9 | 0 | 0.00 |

Figure 5: The screen shot following selection of the file to be analyzed, which allows the data to viewed in the "Data Viewer" panel.



Welcome to RJafroc v0.1.0!

Select Data File: frocData.xlsx

```
Click "Analyze" to generate analysis report.
```

Data type: FROC

Number of modalities: 2

Number of readers: 4

Number of normal cases: 200

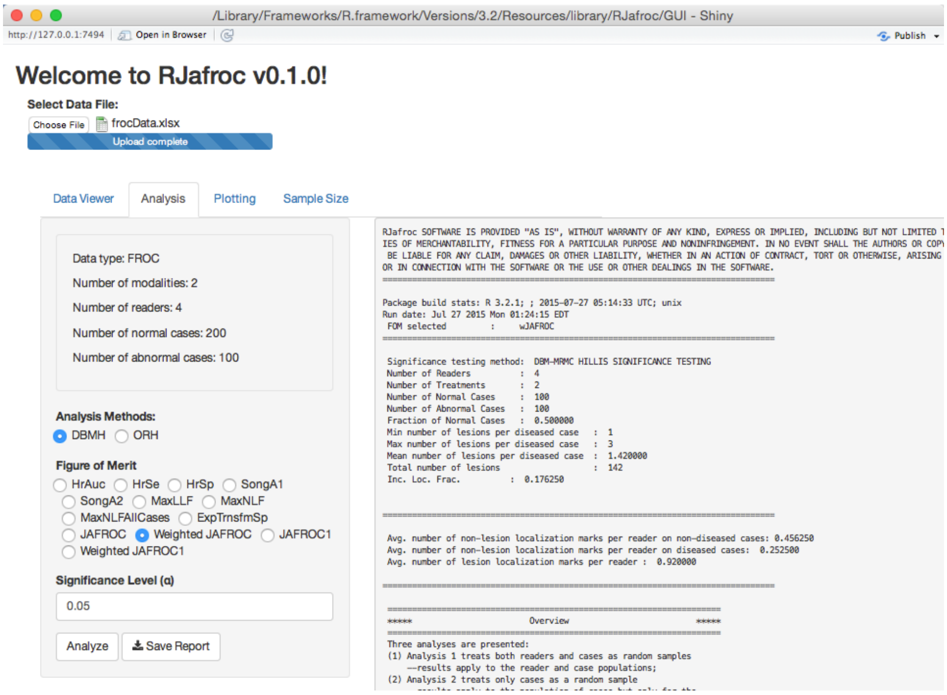
Number of abnormal cases: 100

Analysis Methods:
● DBMH ○ ORH

Figure of Merit
○ HrAuc ○ HrSe ○ HrSp ○ SongA1
○ SongA2 ○ MaxLLF ○ MaxNLF
○ MaxNLFAllCases ○ ExpTrnsfmSp
○ JAFROC ● Weighted JAFROC ○ JAFROC1
○ Weighted JAFROC1

Significance Level (α)
0.05

Figure 6: This screen shows a summary of the dataset, and allows selection of analysis method, the FOM and the significance level of the testing.



Welcome to RJafroc v0.1.0!

Select Data File: frocData.xlsx

Data type: FROC

Number of modalities: 2

Number of readers: 4

Number of normal cases: 200

Number of abnormal cases: 100

Analysis Methods:
● DBMH ○ ORH

Figure of Merit
○ HrAuc ○ HrSe ○ HrSp ○ SongA1
○ SongA2 ○ MaxLLF ○ MaxNLF
○ MaxNLFAllCases ○ ExpTrnsfmSp
○ JAFROC ● Weighted JAFROC ○ JAFROC1
○ Weighted JAFROC1

Significance Level (α)
0.05

```
RJafroc SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED T
IES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPY
 BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING
OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.
================================================================================

Package build stats: R 3.2.1; ; 2015-07-27 05:14:33 UTC; unix
Run date: Jul 27 2015 Mon 01:24:15 EDT
 FOM selected         :     wJAFROC
================================================================================

 Significance testing method:  DBM-MRMC HILLIS SIGNIFICANCE TESTING
 Number of Readers          :  4
 Number of Treatments       :  2
 Number of Normal Cases     :  100
 Number of Abnormal Cases   :  100
 Fraction of Normal Cases   :  0.500000
 Min number of lesions per diseased case   :  1
 Max number of lesions per diseased case   :  3
 Mean number of lesions per diseased case  :  1.420000
 Total number of lesions                   :  142
 Inc. Loc. Frac.          :  0.176250


================================================================================

 Avg. number of non-lesion localization marks per reader on non-diseased cases: 0.456250
 Avg. number of non-lesion localization marks per reader on diseased cases:  0.252500
 Avg. number of lesion localization marks per reader :  0.920000

================================================================================

 ======================================================================
 *****                       Overview                            *****
 ======================================================================
 Three analyses are presented:
 (1) Analysis 1 treats both readers and cases as random samples
     --results apply to the reader and case populations;
 (2) Analysis 2 treats only cases as a random sample
     --results apply to the population of cases but only for the
```

Figure 7: This screen shot shows the results of the analysis. The format closely follows that of Windows JAFROC. Clicking on the “Save” button allows the results to be saved to a text file.

| Software | OR-DBM MRMC | iMRMC | JAFROC | **RJafroc** |
|---|---|---|---|---|
| Data entry | Plain text in specified format | Plain text in specified format | Excel file in JAFROC format | All text and Excel file formats |
| Open Source/Language | No/Fortran/C++ | Yes/Java | No/C++ | Yes/R |
| Cross platform | No | Yes | No | Yes |
| Call from other Languages | No | No | No | Yes |
| ROC curve fitting | Yes | No | No | No |
| Integrated data visualization capability | No | Yes | Yes | Yes |
| Localization paradigms (ROI and FROC) | No | No | Yes | Yes |
| Predicting search paradigm operating characteristics | No | No | Yes | Yes |
| Saving an ROC dataset in a different format | No | No | No | Yes |

Table 10: Software capabilities comparison of available methods of analyzing observer performance data

## 4. Discussion

This paper has covered several topics relevant to assessment of medical imaging systems. These include the choice of data collection paradigm (ROC, FROC or ROI), the choice of figure of merit(the Wilcoxon statistic for ROC data, weighted area under AFROC and several other measures for FROC data, and the trapezoidal area under the ROC' curve for ROI data), significance-testing methods (DBMH and ORH), and sample-size estimation for ROC studies. Data visualization methods have been described for ROC and FROC studies. Future plans call for implementing curve-fitting (as opposed to empirical curves) algorithms that have been described in the literature for ROC data, and methods under development, see below, for FROC data. In our experience statisticians tend to favor the empirical FOMs, as these are least based on what might be considered as "restrictive" assumptions. However, when operating points do not span the ROC x-axis adequately (e.g., they are bunched close to the left edge of the ROC plot), then empirical FOMs become quite dependent on the spread of the points, which can lead to misleading inferences. For this reason it is important to implement curve-fitting procedures (these are available for ROC data in the University of Iowa and University of Chicago website software, but these are not open-source).

A preliminary sample-size method for free-response studies is available on the JAFROC website. The problem is essentially one of determining the JAFROC effect size that would correspond to a particular inferred ROC effect size. Effect sizes are well understood in ROC methodology, since the paradigm dates to the early 1940s (it was originally introduced (Hilden 1991) to measure performance of radar in detecting enemy aircraft). The other FOMs, particularly FROC, are less well understood, as evidenced, in the computer aided detection field, by the widespread usage of the FROC curve to measure performance (we have seen that this is a poor representation of performance, a fact recognized by at least one other FDA researcher familiar with this research area (Popescu 2011)). Attempts to analyze FROC data began in the late 70s (Bunch *et al.* 1978) and some progress was made in the late 80s and early 90s(Chakraborty *et al.* 1986; Chakraborty 1989; Chakraborty and Winter 1990). However, until 2004 (Chakraborty and Berbaum 2004) there was no validated way of analyzing FROC data. To assign a realistic effect size for an FROC figure of merit one needs a model for fitting FROC data that also predicts ROC data. Such a model (the Chakraborty search model) has been introduced (Chakraborty 2006b,a) and a preliminary maximum likelihood estimation method is implemented in the Windows version of JAFROC software. The fits are performed for each reader and modality; for each modality the 3 search-model parameters per reader are averaged, and the average values are used to predict two AFROC and two ROC curves, one per modality. This relates the AFROC area effect size to the ROC area effect size (the former is larger) and permits sample size estimation using the AFROC area as the FOM. For improved reliability, we are currently working on enhancements to the estimation procedure (essentially by imposing a constraint). Another direction for improvement is accommodating the LROC paradigm, currently unsupported by any easily accessible software, but clinically quite important. We also plan to implement the analytical ROI method in a future update.

The choice of data collection paradigm depends on the clinical application. The FROC paradigm is appropriate for the chest lung nodule detection task, but detection of diffuse chest disease (e.g., interstitial lung disease or pneumoconiosis) is appropriately analyzed by the ROC paradigm. Myocardial perfusion imaging, which involves scoring each of three main vascular territories in the heart (Volokh, Liu, and Tsui 2006), is appropriately analyzed by the ROI method. There are also situations where the LROC paradigm is appropriate. The idea is to match the paradigm to the clinical task, and this requires clinical input and as one may expect, “one size does not fit all”. This is the main reason why we have embraced all data collection paradigms in our software. It is our hope that this open source release will stimulate further research, particularly in the extensions to ROC methodology.

## 5. Acknowledgments

We are grateful to Dr. Federica Zanca and Dr. Kevin Berbaum for providing us with datasets. Dr. Hong-Jun Yoon did much of the original Visual C++ programming for the Windows version of JAFROC. Funding from the Department of Health and Human Services, National Institutes of Health (R01-EB005243), supported this work.

## References

Bamber D (1975). “The Area Above the Ordinal Dominance Graph and the Area

Below the Receiver Operating Characteristic Graph.” *Journal of Mathematical Psychology*, **12**(4), 387–415. ISSN 0022-2496. doi:Doi:10.1016/0022-2496(75)90001-2. URL http://www.sciencedirect.com/science/article/B6WK3-4D7JNKG-8D/2/752ed837f02a9523cda7e96258f5516c.

Barnes G, Sabbagth E, Chakraborty D, Nath P, Luna R, Sanders C, Fraser R (1989). “A Comparison of Dual-energy Digital Radiography and Screen-film Imaging in the Detection of Subtle Interstitial Pulmonary Disease.” *Invest Radiol.*, **24**(8), 585–591.

Beiden SV, Wagner RF, Campbell G (2000). “Components-of Variance Models and Multiple-Bootstrap Experiments: An Alternative Method for Random-Effects, Receiver Operating Characteristic Analysis.” *Academic Radiology*, **7**(5), 341–349.

Bunch PC, Hamilton JF, Sanderson GK, Simmons AH (1978). “A Free-Response Approach to the Measurement and Characterization of Radiographic-Observer Performance.” *J of Appl Photogr. Eng.*, **4**(4), 166–171.

Chakraborty D, Breatnach E, Yester M, Soto B, Barnes G, Fraser R (1986). “Digital and Conventional Chest Imaging: A Modified ROC Study of Observer Performance Using Simulated Nodules.” *Radiology*, **158**, 35–39.

Chakraborty DP (1989). “Maximum Likelihood Analysis of Free-response Receiver Operating Characteristic (FROC) Data.” *Med. Phys.*, **16**(4), 561–568.

Chakraborty DP (2002). “Statistical Power in Observer Performance Studies: a Comparison of the ROC and Free-response Methods in Tasks Involving Localization.” *Acad. Radiol.*, **9**(2), 147–156.

Chakraborty DP (2006a). “ROC Curves Predicted by a Model of Visual Search.” *Phys. Med. Biol.*, **51**, 3463–3482.

Chakraborty DP (2006b). “A Search Model and Figure of Merit for Observer Data Acquired According to the Free-response Paradigm.” *Phys. Med. Biol.*, **51**, 3449–3462.

Chakraborty DP (2008). “Validation and Statistical Power Comparison of Methods for Analyzing Free-response Observer Performance Studies.” *Acad Radiol*, **15**(12), 1554–1566. URL http://www.sciencedirect.com/science/article/B75BK-4TW6D0R-9/2/8f59ae9ff4ba7d2aa596076694b7de09.

Chakraborty DP (2012). “Measuring Agreement Between Ratings Interpretations and Binary Clinical Interpretations of Images: a Simulation Study of Methods for Quantifying the Clinical Relevance of an Observer Performance Paradigm.” *Phys. Med. Biol.*, **57**, 2873–2904. URL http://stacks.iop.org/0031-9155/57/2873.

Chakraborty DP (2013). “A Brief History of Free-Response Receiver Operating Characteristic Paradigm Data Analysis.” *Academic Radiology*, **20**(7), 915–919. ISSN 1076-6332. doi:http://dx.doi.org/10.1016/j.acra.2013.03.001. URL http://www.sciencedirect.com/science/article/pii/S1076633213001104.

Chakraborty DP, Berbaum KS (2004). “Observer Studies Involving Detection and Localization: Modeling, Analysis and Validation.” *Med Phys*, **31**(8), 2313–2330.

Chakraborty DP, Winter LHL (1990). "Free-Response Methodology: Alternate Analysis and a New Observer-Performance Experiment." *Radiology*, **174**, 873–881.

Chakraborty DP, Zhai X (2014). *JAFROC Software, Version 4.2.1.* URL http://www.devchakraborty.com/.

DeLong ER, DeLong DM, Clarke-Pearson DL (1988). "Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach." *Biometrics*, **44**, 837–845.

Dorfman D, Alf E (1968). "Maximum Likelihood Estimation of Parameters of Signal Detection Theory- a Direct Solution." *Psychometrika*, **33**(1), 117–124.

Dorfman D, Alf E (1969). "Maximum-Likelihood Estimation of Parameters of Signal-Detection Theory and Determination of Confidence Intervals - Rating-Method Data." *Journal of Mathematical Psychology*, **6**, 487–496.

Dorfman D, Berbaum K (1986). "RSCORE-J: Pooled Rating-method Data: a Computer Program for Analyzing Pooled ROC Curves." *Behavior Research Methods, Instruments, & Computers*, **18**(5), 452–462.

Dorfman D, Berbaum K (2000). "A Contaminated Binormal Model for ROC Data: Part II. a Formal Model." *Acad Radiol.*, **7**(6), 427–37.

Dorfman D, Berbaum K, Metz C (1992). "ROC Characteristic Rating Analysis: Generalization to the Population of Readers and Patients with the Jackknife Method." *Invest. Radiol.*, **27**(9), 723–731.

Dorfman D, Berbaum K, Metz C, Lenth R, Hanley J, Abu Dagga H (1997). "Proper Receiving Operating Characteristic Analysis: The Bigamma model." *Acad. Radiol.*, **4**(2), 138–149.

Efron B, Tibshirani RJ (1993). *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton.

Egan J, Greenburg G, Schulman A (1961). "Operating Characteristics, Signal Detectability and the Method of Free Response." *J Acoust Soc. Am.*, **33**, 993–1007.

Foucher Y (2013). *Construction of Diagnostic or Prognostic Scoring System and Internal Validation of Its Discriminative Capacities Based on ROC Curve and 0.633 Boostrap Resampling.* URL http://CRAN.R-project.org/package=ROC632.

Foucher Y, Trebern-Launay K, Lorent M (2013). ***ROCt**: Time-dependent ROC Curve Estimation and Adaptation to the Relative Survival Context.* R package version 0.8, URL http://CRAN.R-project.org/package=ROCt.

Gallas BD (2006). "One-Shot Estimate of MRMC Variance: AUC." *Academic Radiology*, **13**(3), 353–362. ISSN 1076-6332. URL http://www.sciencedirect.com/science/article/B75BK-4J915RW-C/2/11387ee9975e90e6ae25bef247c817de.

Gallas BD, Bandos A, Samuelson FW, Wagner RF (2009). "A Framework for Random-Effects ROC Analysis: Biases with the Bootstrap and Other Variance Estimators." *Communications in Statistics - Theory and Methods*, **38**(15), 2586 – 2603. ISSN 0361-0926. URL http://www.informaworld.com/10.1080/03610920802610084.

Gallas BD, He X, Pathare R (2015). *iMRMC Software, Version 2.7.* URL https://code.google.com/p/imrmc/.

Hanley JA, McNeil BJ (1982). "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve." *Radiology*, **143**(1), 29–36. URL http://radiology.rsnajnls.org/cgi/content/abstract/143/1/29.

Hilden J (1991). "The Area Under the ROC Curve and Its Competitors." *Medical Decision Making*, **11**, 95 –101.

Hillis SL (2007). "A Comparison of Denominator Degrees of Freedom Methods for Multiple Observer ROC Studies." *Statistics in Medicine*, **26**, 596–619.

Hillis SL, Berbaum KS, Metz CE (2008). "Recent Developments in the Dorfman-Berbaum-Metz Procedure for Multireader ROC Study Analysis." *Academic radiology*, **15**(5), 647–661.

Hillis SL, Obuchowski NA, Berbaum KS (2011). "Power Estimation for Multireader ROC Methods: An Updated and Unified Approach." *Academic Radiology*, **18**(2), 129–142. ISSN 1076-6332. doi:DOI:10.1016/j.acra.2010.09.007. URL http://www.sciencedirect.com/science/article/B75BK-51XH2HN-2/2/c1b2bd3ca6c51d50f99dc36327c1362b.

Hillis SL, Schartz KM, Berbaum KS (2015). *OR-DBM MRMC Software, Version 2.5.* URL http://perception.radiology.uiowa.edu/.

Ishwaran H, Gatsonis CA (2000). "A General Class of Hierarchical Ordinal Regression Models with Applications to Correlated ROC Analysis." *The Canadian Journal of Statistics*, **28**(4), 731–750.

Jiang Y, Metz CE, Nishikawa RM (1996). "A Receiver Operating Characteristic Partial Area Index for Highly Sensitive Diagnostic Tests." *Radiology*, **201**(3), 745–750. ISSN 0033-8338.

Kundel H, Berbaum K, Dorfman D, Gur D, Metz CE, Swensson RG (2008). "Receiver Operating Characteristic Analysis in Medical Imaging (ICRU Report 79)." *Report*, International Commission on Radiation Units & Measurments.

Kundel HL, Nodine CF, Conant EF, Weinstein SP (2007). "Holistic Component of Image Perception in Mammogram Interpretation: Gaze-tracking Study." *Radiology*, **242**(2), 396–402.

Mann HB, Whitney DR (1947). "On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other." *The Annals of Mathematical Statistics*, **18**, 50âĹŠ60.

Metz C (1978). "Basic Principles of ROC Analysis." *Seminars in Nuclear Medicine*, **8**(4), 283–298.

Metz C (1989). "Some Practical Issues of Experimental Design and Data Analysis in Radiological ROC studies." *Investigative Radiology*, **24**, 234–245.

Metz C (2008). "ROC Analysis in Medical Imaging: a Tutorial Review of the Literature." *Radiological Physics and Technology*, **1**(1), 2–12. ISSN 1865-0333. doi:10.1007/s12194-007-0002-1. URL http://dx.doi.org/10.1007/s12194-007-0002-1.

Metz C, Pan X (1999). "Proper Binormal ROC Curves: Theory and Maximum-Likelihood Estimation." *J Math Psychol*, **43**(1), 1–33.

Metz CE (1986). "ROC Methodology in Radiologic Imaging." *Investigative Radiology*, **21**(9), 720–733. ISSN 0020-9996. URL http://journals.lww.com/investigativeradiology/Fulltext/1986/09000/ROC_Methodology_in_Radiologic_Imaging.9.aspx.

Metz CE, Herman BA, Roe CE (1998a). "Statistical Comparison of Two ROC-curve Estimates Obtained from Partially-paired Datasets." *Med Decis Making*, **18**(1), 110–121. doi:doi:10.1177/0272989X9801800118.

Metz CE, Herman BA, Shen JH (1998b). "Maximum Likelihood Estimation of Receiver Operating Characteristic (ROC) Curves from Continuously-distributed Data." *Statistics in medicine*, **17**(9), 1033–1053.

Miller H (1969). "The FROC Curve: a Representation of the Observer's Performance for the Method of Free Response." *The Journal of the Acoustical Society of America*, **46**(6(2)), 1473–1476.

Niklason LT, Hickey NM, Chakraborty DP, Sabbagh EA, Yester MV, Fraser RG, Barnes GT (1986). "Simulated Pulmonary Nodules: Detection with Dual-energy Digital Versus Conventional Radiography." *Radiology*, **160**, 589–593.

Obuchowski N (2009). "Reducing the Number of Reader Interpretations in MRMC Studies." *Acad Radiol*, **16**, 209–217.

Obuchowski NA (1997). "Nonparametric Analysis of Clustered ROC Curve Data." *Biometrics*, **53**, 567–578.

Obuchowski NA, Beiden SV, Berbaum KS, Hillis SL, Ishwaran H, Song H, Wagner RF (2004). "Multireader, Multicase Receiver Operating Characteristic Analysis: An Empirical Comparison of Five Methods." *Acad. Radiol.*, **11**(9), 980–995.

Obuchowski NA, Lieber ML, Powell KA (2000). "Data Analysis for Detection and Localization of Multiple Abnormalities with Application to Mammography." *Acad. Radiol.*, **7**(7), 516–525.

Obuchowski NA, Mazzone PJ, Dachman AH (2010). "Bias, Underestimation of Risk, and Loss of Statistical Power in Patient-level Analyses of Lesion Detection." *Eur Radiol*, **20**, 584–594.

Obuchowski NA, Rockette H (1995). "Hypothesis Testing of the Diagnostic Accuracy for Multiple Diagnostic Tests: An ANOVA Approach with Dependent Observations." *Communications in Statistics: Simulation and Computation*, **24**, 285–308.

Pan X, Metz C (1997). "The Proper Binormal Model: Parametric Receiver Operating Characteristic Curve Estimation with Degenerate Data." *Acad. Radiol.*, **4**(5), 380–9.

Pepe MS (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford Statistical Science Series. Oxford University Press, New York.

Pesce LL, Metz CE (2007). "Reliable and Computationally Efficient Maximum-Likelihood Estimation of Proper Binormal ROC Curves." *Acad Radiol*, **14**(7), 814–829.

Pesce LL, Papaioannou J, Metz CE (2011). *Java GUI ROC Software, Version 1.0.3.* URL http://metz-roc.uchicago.edu/MetzROC/software/.

Pisano E, Gatsonis C, Hendrick E, Yaffe M, Baum J, Acharyya S, Conant E, Fajardo L, Bassett L, D'Orsi C, Jong R, Rebner M (2005). "Diagnostic Performance of Digital Versus Film Mammography for Breast-cancer Screening." *N Engl J Med*, **353**(17), 1–11.

Popescu LM (2011). "Nonparametric Signal Detectability Evaluation Using an Exponential Transformation of the FROC Curve." *Medical Physics*, **38**(10), 5690–5702.

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M (2011). "**pROC**: an Open-source Package for R and S to Analyze and Compare ROC Curves." *BMC bioinformatics*, **12**(1), 77.

Roe CA, Metz C (1997). "Dorfman-Berbaum-Metz Method for Statistical Analysis of Multi-reader, Multimodality Receiver Operating Characteristic Data: Validation with Computer Simulation." *Acad Radiol*, **4**, 298–303.

Sing T, Sander O, Beerenwinkel N, Lengauer T (2005). "**ROCR**: Visualizing Classifier Performance in R." *Bioinformatics*, **21**(20), 3940–3941. doi:10.1093/bioinformatics/bti623.

Song H (1997). "Analysis of Correlated Roc Areas in Diagnostic Testing." *Biometrics*, **53**, 370–382.

Song T, Bandos AI, Rockette HE, Gur D (2008). "On Comparing Methods for Discriminating Between Actually Negative and Actually Positive Subjects with Froc Type Data." *Medical Physics*, **35**(4), 1547–1558.

Starr S, Metz C, Lusted L (1977). "Comments on Generalization of Receiver Operating Characteristic Analysis to Detection and Localization Tasks." *Phys. Med. Biol.*, **22**, 376–379.

Starr SJ, Metz CE, Lusted LB, Goodenough DJ (1975). "Visual Detection and Localization of Radiographic Images." *Radiology*, **116**, 533–538.

Swensson R, Judy P (1981). "Detection of Noisy Visual Targets: Models for the Effects of Spatial Uncertainty and Signal-to-noise Ratio." *Perception & Psychophysics*, **29**(6), 521–534.

Swensson RG (1996). "Unified Measurement of Observer Performance in Detecting and Localizing Target Objects on Images." *Med. Phys.*, **23**(10), 1709 –1725.

Toledano A, Gatsonis C (1996). "Ordinal Regression Methodology for ROC Curves Derived from Correlated Data." *Stat Med*, **15**(16), 1807–1826.

Toledano AY (2003). "Three Methods for Analyzing Correlated ROC Curves: a Comparison in Real Data Sets." *Statistics in Medicine*, **22**(18), 2919–33.

Van Dyke C, White R, Obuchowski N, Geisinger M, Lorig R, Meziane M (1993). "Cine MRI in the Diagnosis of Thoracic Aortic Dissection." *79th RSNA Meetings*.

Volokh L, Liu C, Tsui BM (2006). "Exploring FROC Paradigm: Initial Experience with Clinical Applications." In *Medical Imaging*, pp. 61460D–61460D. International Society for Optics and Photonics.

Wagner R, Beiden S, Campbell G, Metz C, Sacks W (2002). "Assessment of Medical Imaging and Computer-assist Systems: Lessons from Recent Experience." *Academic Radiology*, **9**(11), 1264 –1277.

Wagner RF, Metz CE, Campbell G (2007). "Assessment of Medical Imaging Systems and Computer Aids: A Tutorial Review." *Acad Radiol*, **14**(6), 723–748.

Wang C (2012). *Non-parametric Estimation of ROC Curves Without Gold Standard Test.* URL `http://CRAN.R-project.org/package=ROCwoGS`.

Wickham H (2009). ***ggplot2**: Elegant Graphics for Data Analysis.* Springer-Verlag.

Wilcoxon F (1945). "Individual Comparison by Ranking Methods." *Biometrics*, **1**, 80–83.

Youden W (1950). "Index for Rating Diagnostic Tests." *Cancer*, **3**, 32–35.

Yousef W, Wagner R, Loew M (2005). "The Partial Area Under the ROC Curve: Its Properties and Nonparametric Estimation for Assessing Classifier Performance."

Zanca F, Jacobs J, Van Ongeval C, Claus F, Celis V, Geniets C, Provost V, Pauwels H, Marchal G, Bosmans H (2009). "Evaluation of Clinical Image Processing Algorithms Used in Digital Mammography." *Medical Physics*, **36**(3), 765–775.

# A. Appendices

## A.1. Other free-response FOMs implemented in RJafroc

Free-response data can be used to infer maximum sensitivity and specificity, corresponding to the highest operating point on the ROC curve, excluding the trivial point at $(1, 1)$. These are defined by

$$\begin{aligned}\widehat{\theta}_{ij}^{ISe} &= \tfrac{1}{K_2}\sum_{k_2=1}^{K_2} \phi\left(max\left(r_{ijk_2 2**}\right)\right)\\ \widehat{\theta}_{ij}^{ISp} &= 1 - \tfrac{1}{K_1}\sum_{k_1=1}^{K_1} \phi\left(max\left(r_{ijk_1 1*1}\right)\right)\end{aligned}$$

The JAFROC FOM is defined as the probability that lesions are rated higher than the highest noise on *normal* images:

$$\widehat{\theta}_{ij}^{JAFROC} = \frac{1}{N_2 K_1}\sum_{k_2=1}^{K_2}\sum_{l_2=1}^{N_{k_2 2}}\sum_{k_1=1}^{K_1} \psi\left(max(r_{ijk_1 1*1}), r_{ijk_2 2l_2 2}\right) \tag{A.1}$$

The corresponding JAFROC1 FOM, which includes the highest noise on abnormal images, is defined by

$$\begin{aligned}\widehat{\theta}_{ij}^{JAFROC1} = \frac{1}{N_2\left(K_1 + K_2\right)}\sum_{k_2=1}^{K_2}\sum_{l_2=1}^{N_{k_2 2}}\Bigg[&\sum_{k_1=1}^{K_1} \psi\left(max(r_{ijk_1 1*1}), r_{ijk_2 2l_2 2}\right)\\ &+\sum_{k_2'=1}^{K_2} \psi\left(max(r_{ijk_2' 2*1}), r_{ijk_2 2l_2 2}\right)\Bigg]\end{aligned} \tag{A.2}$$

The maximum LLF FOM is defined by

$$\widehat{\theta}_{ij}^{\max LLF} = \frac{\sum\limits_{k_2=1}^{K_2}\sum\limits_{l_2=1}^{N_{k_2 2}} \phi\left(r_{ijk_2 2l_2 2}\right)}{N_2} \tag{A.3}$$

Here $\phi\left(x\right) = 1$ is $x$ is finite and $\phi\left(-\infty\right) = 0$. The maximum NLF FOM is defined by

$$\widehat{\theta}_{ij}^{\max NLF} = \frac{\sum\limits_{k_1=1}^{K_1}\sum\limits_{k_2=1}^{K_2}\sum\limits_{l_i=1}^{N_{k_t t1}} \phi\left(r_{ijk_t tl_1 1}\right)}{K_1 + K_2} \tag{A.4}$$

An exponentially transformed specificity FOM (Popescu 2011) is defined by:

$$\widehat{\theta}_{ij}^{IExpTrnsSp} = \exp\left( -\frac{\sum\limits_{k_1=1}^{K_1} \sum\limits_{l_i=1}^{N_{k_1 1 1}} \phi\left(r_{ijk_1 1 l_1 1}\right)}{K_1} \right) \tag{A.5}$$

These are summarized in Table A.1:

Table A.1

| Paradigm | Description of FOM | Symbol | Comments |
|---|---|---|---|
| FROC | Highest rating inferred sensitivity | $\widehat{\theta}_{ij}^{ISe}$ | Case-level inferred sensitivity |
| | Highest rating inferred specificity | $\widehat{\theta}_{ij}^{ISp}$ | Case-level inferred specificity |
| | Exponentially transformed specificity | $\widehat{\theta}_{ij}^{IExpTrnsSp}$ | Popescu suggestion |
| | JAFROC | $\widehat{\theta}_{ij}^{JAFROC}$ | Does not use weighting |
| | JAFROC1 | $\widehat{\theta}_{ij}^{JAFROC1}$ | Does not use weighting |
| | Maximum ordinate of FROC | $\widehat{\theta}_{ij}^{maxLLF}$ | Lesion-level ”sensitivity” |
| | Maximum abscissa of FROC | $\widehat{\theta}_{ij}^{maxNLF}$ | Lesion-level ”inverse specificity”, lower values preferred |

## A.2. Special cases of DBMH analysis

*Fixed-reader random-case (FRRC) analysis*

When readers are treated as a fixed effect, the appropriate $f$ statistic for testing the null hypothesis is

$$F_{DBM|R} = \frac{MS_Y\left(T\right)}{MS_Y\left(TC\right)} \tag{A.6}$$

This is distributed as an $f$ statistic with $ndf = I - 1$, and $ddf = (I - 1)(K - 1)$:

$$F_{DBM|R} \sim F_{I-1,(I-1)(K-1)} \tag{A.7}$$

The critical value of the statistic is $F_{1-\alpha;I-1,(I-1)(K-1)}$ which is that value such that fraction $(1 - \alpha)$ of the distribution lies to the left of the critical value. The null hypothesis is rejected if the observed value of the $f$ statistic exceeds the critical value:

$$F_{DBM|R} > F_{1-\alpha;I-1,(I-1)(K-1)} \tag{A.8}$$

The $p$-value of the test is the probability that a random sample from the distribution exceeds the observed value:

$$p = \mathrm{P}\left(F > F_{DBM|R} | F \sim F_{I-1,(I-1)(K-1)}\right) \tag{A.9}$$

The $(1-\alpha)$ confidence interval is given by:

$$CI_{1-\alpha} = \left(\widehat{\theta}_{i\bullet} - \widehat{\theta}_{i'\bullet}\right) \pm t_{\alpha/2;(I-1)(K-1)} \sqrt{\frac{2}{JK} MS_Y\left(TC\right)} \tag{A.10}$$

*Random-reader fixed case (RRFC) analysis*

When cases are treated as a fixed effect, the appropriate $f$ statistic for testing the null hypothesis is

$$F_{DBM|C} = \frac{MS_Y\left(T\right)}{MS_Y\left(TR\right)} \tag{A.11}$$

This is distributed as an $f$ statistic with $ndf = I-1$, and $ddf = (I-1)(J-1)$:

$$F_{DBM|C} \sim F_{I-1,(I-1)(J-1)} \tag{A.12}$$

The critical value of the statistic is $F_{1-\alpha;I-1,(I-1)(J-1)}$ which is that value such that fraction $(1-\alpha)$ of the distribution lies to the left of the critical value. The null hypothesis is rejected if the observed value of the $f$ statistic exceeds the critical value:

$$F_{DBM|C} > F_{1-\alpha;I-1,(I-1)(J-1)} \tag{A.13}$$

The $p$-value of the test is the probability that a random sample from the distribution exceeds the observed value:

$$p = \mathrm{P}\left(F > F_{DBM|C} | F \sim F_{I-1,(I-1)(J-1)}\right) \tag{A.14}$$

The $(1-\alpha)$ confidence interval is given by:

$$CI_{1-\alpha} = \left(\widehat{\theta}_{i\bullet} - \widehat{\theta}_{i'\bullet}\right) \pm t_{\alpha/2;(I-1)(J-1)} \sqrt{\frac{2}{JK} MS_Y\left(TR\right)} \tag{A.15}$$

## A.3. Special cases of ORH analysis

*Fixed-reader random-case (FRRC) analysis*

When readers are treated as a fixed effect, the appropriate $f$ statistic for testing the null hypothesis is

$$F_{OR|R} = \frac{MS\left(T\right)}{\left[\widehat{Var} - \widehat{Cov}_1 + (J-1)\,H\left(\widehat{Cov}_2 - \widehat{Cov}_3\right)\right]} \tag{A.16}$$

This is distributed as an $f$ statistic with $ndf = I-1$, and $ddf = \infty$, or equivalently a chi-square distribution with I-1 degrees of freedom:

$$F_{OR|R} \sim F_{I-1,\infty} = \chi^2_{I-1} \tag{A.17}$$

The critical value of the statistic is $F_{1-\alpha;I-1,\infty} = \chi^2_{1-\alpha;I-1,}$ which is that value such that fraction $(1-\alpha)$ of the distribution lies to the left of the critical value. The null hypothesis is rejected if the observed value of the $f$ statistic exceeds the critical value:

$$F_{OR|R} > F_{1-\alpha;I-1,\infty} = \chi^2_{1-\alpha;I-1,} \tag{A.18}$$

The $p$ value of the test is the probability that a random sample from the distribution exceeds the observed value:

$$p = \mathrm{P}\left(F > F_{OR|R} | F \sim F_{I-1,\infty}\right) \tag{A.19}$$

The $(1-\alpha)$ confidence interval is given by:

$$CI_{1-\alpha} = \left(\widehat{\theta}_{i\bullet} - \widehat{\theta}_{i'\bullet}\right) \pm t_{\alpha/2;\infty}\sqrt{\frac{2}{J}\left[\widehat{Var} - \widehat{Cov}_1 + (J-1)\,H\left(\widehat{Cov}_2 - \widehat{Cov}_3\right)\right]} \tag{A.20}$$

*Random-reader fixed case (RRFC) analysis*

When cases are treated as a fixed effect, the appropriate $f$ statistic for testing the null hypothesis is:

$$F_{OR|C} = \frac{MS\,(T)}{MS\,(TR)} \tag{A.21}$$

This is distributed as:

$$F_{OR|C} \sim F_{I-1,(I-1)(J-1)} \tag{A.22}$$

The critical value of the statistic is $F_{1-\alpha;I-1,(I-1)(J-1)}$ which is that value such that fraction $(1-\alpha)$ of the distribution lies to the left of the critical value. The null hypothesis is rejected if the observed value of the $f$ statistic exceeds the critical value:

$$F_{DBM|C} > F_{1-\alpha;I-1,(I-1)(J-1)} \tag{A.23}$$

The $p$ value of the test is the probability that a random sample from the distribution exceeds the observed value:

$$p = \mathrm{P}\left(F > F_{DBM|C} | F \sim F_{I-1,(I-1)(J-1)}\right) \tag{A.24}$$

The $(1-\alpha)$ confidence interval is given by:

$$CI_{1-\alpha} = \left(\widehat{\theta}_{i\bullet} - \widehat{\theta}_{i'\bullet}\right) \pm t_{\alpha/2;(I-1)(K-1)}\sqrt{\frac{2}{J}MS\,(TR)} \tag{A.25}$$

## A.4. Details of ROI simulator

Since it is based on the Roe-Metz simulator for ROC data, we begin by describing the ROC data simulator for MRMC studies. For each modality, it consists of two unit variance distributions separated by an amount that determines AUC in that modality. The readers and cases are modeled by random samples and there is an error term that depends on treatments, readers and cases. The Roe and Metz model is (Roe and Metz 1997):

$$Z_{ijk_tt} = \mu_t + \tau_{it} + C_{k_tt} + R_{jt} + (\tau C)_{ik_tt} + (\tau R)_{ijt} + (RC)_{jk_tt} + \varepsilon_{ijk_tt} \tag{A.26}$$

The fixed effects in the simulator are described by

$$\begin{aligned} &\mu_1 = 0; \mu_2 = \mu \\ &\tau_{i,1} = 0; \tau_{1,2} = 0; \tau_{2,1} = \tau \end{aligned} \tag{A.27}$$

The random effects are described by

$$\begin{aligned} &C_{k_tt} \sim N\left(0, \sigma_C^2\right) \\ &R_{jt} \sim N\left(0, \sigma_R^2\right) \\ &(\tau C)_{ik_tt} \sim N\left(0, \sigma_{\tau C}^2\right) \\ &(\tau R)_{ijt} \sim N\left(0, \sigma_{\tau R}^2\right) \\ &(RC)_{jk_tt} \sim N\left(0, \sigma_{RC}^2\right) \\ &\varepsilon_{ijk_tt} \sim N\left(0, \sigma_\varepsilon^2\right) \end{aligned} \tag{A.28}$$

To preserve the unit variance character of the model, the following constraint is applied:

$$\sigma_C^2 + \sigma_{\tau C}^2 + \sigma_{RC}^2 + \sigma_\varepsilon^2 = 1 \tag{A.29}$$

Since ROI data is a special case of FROC data, we denote the ROI rating by $r_{ijk_ttl_ss}$ where on$l_1 = 1, 2, \ldots, Q$ non-diseased cases, where $Q$ is the number of ROIs (or “quadrants”) on every case, and on diseased cases $l_2 = 1, 2, ..., q_{k_2}$, where $q_{k_2}$ is the number of diseased ROIs on diseased case $k_2 2$, and $l_1 = 1, 2, \ldots, Q - q_{k_2}$ on diseased case $k_2 2$.

The ROI model is defined by:

$$\begin{aligned} Z_{ijk_tt} = \mu_t + \tau_{it} + C_{k_tt} + R_{jt} + (\tau C)_{ik_tt} + (\tau R)_{ijt} + (RC)_{jk_tt} \\ + (CL)_{k_ttl_ss} + (\tau CL)_{ik_ttl_ss} + (RCL)_{jk_ttl_ss} + \varepsilon_{ijk_ttl_ss} \end{aligned} \tag{A.30}$$

The idea is to split up each term containing the case factor into two terms, one containing the case factor, and the other an additional location factor L (for location) with levels defined by , such that the net case variance is unaltered. The following two terms do not contain the case factor and hence do not need to be split.

$$\begin{aligned} &R_{jt} \sim N\left(0, \sigma_R^2\right) \\ &(\tau R)_{ijt} \sim N\left(0, \sigma_{\tau R}^2\right) \end{aligned}$$

The following term containing only the case factor is split up as follows [the term () controls the correlation between the samples from the different locations on the same case]:

$$\begin{aligned} &C_{k_t t} \sim N\left(0, \rho_C \sigma_C^2\right) \\ &(CL)_{k_t t l_s s} \sim N\left(0, (1-\rho_C)\sigma_C^2\right) \end{aligned}$$

Likewise, the treatment-case factor is split up as follows:

$$\begin{aligned} &\tau C_{i k_t t} \sim N\left(0, \rho_{\tau C} \sigma_{\tau C}^2\right) \\ &(\tau CL)_{i k_t t l_s s} \sim N\left(0, (1-\rho_{\tau C})\sigma_{\tau C}^2\right) \end{aligned}$$

The reader-case factor is split up as follows:

$$\begin{aligned} &RC_{j k_t t} \sim N\left(0, \rho_{RC} \sigma_{RC}^2\right) \\ &(RCL)_{j k_t t l_s s} \sim N\left(0, (1-\rho_{RC})\sigma_{RC}^2\right) \end{aligned}$$

Finally, the error term is split up as follows:

$$\begin{aligned} &\varepsilon_{i j k_t t} \sim N\left(0, \rho_\varepsilon \sigma_\varepsilon^2\right) \\ &(\varepsilon L)_{i j k_t t l_s s} \sim N\left(0, (1-\rho_\varepsilon)\, \sigma_\varepsilon^2\right) \end{aligned}$$

For the simulated data the following values, selected from Table 1 in Roe and Metz (1997), were used:

$$\begin{aligned} &\sigma_R^2 = 0.2; \sigma_{\tau R}^2 = 0.005; \\ &\sigma_C^2 = 0.7; \sigma_{\tau C}^2 = 0.05; \sigma_{RC}^2 = 0.2; \sigma_\varepsilon^2 = 0.05; \end{aligned}$$

The correlation parameters were set as follows:

$$\rho_C = 0.1; \rho_{RC} = 0.1; \rho_{\tau C} = 0.9; \rho_\varepsilon = 0.9;$$

## Affiliation:

Dev P. Chakraborty
Professor
Department of Radiology
University of Pittsburgh
FARP Bldg, Room 212
3362 Fifth Ave
Pittsburgh, PA 15213
United States of America
Telephone: +01/412/480-7318
E-mail: dpc10ster@gmail.edu
URL: http://www.devchakraborty.com